

10/649,100

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日                      2 0 0 3 年    1 月 3 1 日  
Date of Application:

出 願 番 号                      特 願 2 0 0 3 - 0 2 5 0 7 4  
Application Number:  
[ST. 10/C]:                      [ J P 2 0 0 3 - 0 2 5 0 7 4 ]

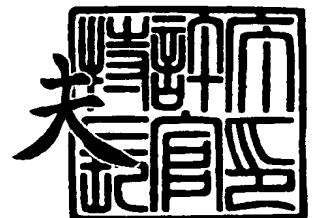
出      願      人                      株式会社日立製作所  
Applicant(s):



2 0 0 3 年    8 月 1 3 日

特許庁長官  
Commissioner,  
Japan Patent Office

今 井 康 夫



出証番号    出証特 2 0 0 3 - 3 0 6 5 1 6 1



【書類名】 特許願

【整理番号】 HI020901

【提出日】 平成15年 1月31日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 3/06

【発明者】

【住所又は居所】 神奈川県横浜市戸塚区戸塚町 5 0 3 0 番地 株式会社日立製作所ソフトウェア事業部内

【氏名】 藤本 修二

【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社日立製作所

【代理人】

【識別番号】 100071283

【弁理士】

【氏名又は名称】 一色 健輔

【選任した代理人】

【識別番号】 100084906

【弁理士】

【氏名又は名称】 原島 典孝

【選任した代理人】

【識別番号】 100098523

【弁理士】

【氏名又は名称】 黒川 恵

【選任した代理人】

【識別番号】 100112748

【弁理士】

【氏名又は名称】 吉田 浩二

## 【選任した代理人】

【識別番号】 100110009

【弁理士】

【氏名又は名称】 青木 康

## 【手数料の表示】

【予納台帳番号】 011785

【納付金額】 21,000円

## 【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 記憶デバイス制御装置、及びプログラム

【特許請求の範囲】

【請求項 1】 情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応する I/O 要求を出力する I/O プロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、

前記各チャネル制御部はフェイルオーバを行うためのグループに類別されている記憶デバイス制御装置であって、

前記各チャネル制御部により更新される前記フェイルオーバ時に引き継がれるデータを、当該チャネル制御部と同一の前記グループに類別されている前記各チャネル制御部が共通にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域である共有ボリュームに記憶する手段を備えることを特徴とする記憶デバイス制御装置。

【請求項 2】 情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応する I/O 要求を出力する I/O プロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、

前記各チャネル制御部はフェイルオーバを行うためのグループに類別されている記憶デバイス制御装置であって、

前記各チャネル制御部により更新される前記フェイルオーバ時に引き継がれるデータを、前記各チャネル制御部が共通にアクセス可能な前記記憶デバイス制御装置が備える共有メモリに記憶する手段を備えることを特徴とする記憶デバイス制御装置。

【請求項 3】 情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応する I/O 要求を出力する I/O プロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、

前記各チャネル制御部はフェイルオーバを行うためのグループに類別されてい

る記憶デバイス制御装置であって、

前記各チャンネル制御部により更新される前記フェイルオーバー時に引き継がれるデータを、当該チャンネル制御部と同一の前記グループに類別されている他の前記チャンネル制御部に、前記各チャンネル制御部を相互に接続するネットワークを介して送信する手段を備えることを特徴とする記憶デバイス制御装置。

【請求項 4】 請求項 1 又は請求項 2 に記載の記憶デバイス制御装置において、

前記各チャンネル制御部には、当該チャンネル制御部が個別にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域であるローカルボリュームがそれぞれ割り当てられており、

前記データを更新した前記チャンネル制御部と同一の前記グループに類別されている他の前記チャンネル制御部の前記ローカルボリュームにも前記データを記憶する手段をさらに備えることを特徴とする記憶デバイス制御装置。

【請求項 5】 請求項 1 又は請求項 2 に記載の記憶デバイス制御装置において、

前記各チャンネル制御部には、当該チャンネル制御部が個別にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域であるローカルボリュームがそれぞれ割り当てられており、

前記データを更新した前記チャンネル制御部と同一の前記グループに類別されている他の前記チャンネル制御部の前記ローカルボリュームにも前記データを記憶する手段と、

前記データの参照先が記録された引き継ぎデータ参照テーブルと、

前記引き継ぎデータ参照テーブルに記録された前記データの参照先に基づき、前記共有ボリューム、前記共有メモリ、又は前記ローカルボリュームのいずれかから前記データを読み出す手段と

をさらに備えることを特徴とする記憶デバイス制御装置。

【請求項 6】 請求項 3 に記載の記憶デバイス制御装置において、

前記データが前記記憶デバイス制御装置内の全ての前記チャンネル制御部に共通に参照されるデータである場合には、当該データを、前記ネットワークを介して

、前記記憶デバイス制御装置内の全ての前記チャンネル制御部に送信する手段を備えることを特徴とする記憶デバイス制御装置。

【請求項 7】 請求項 1 に記載の記憶デバイス制御装置において、

前記データが前記記憶デバイス制御装置内の全ての前記チャンネル制御部に共通に参照されるデータである場合には、当該データを、前記記憶デバイス制御装置内の全ての前記チャンネル制御部が共通にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域である第 2 の共有ボリュームに記憶する手段を備えることを特徴とする記憶デバイス制御装置。

【請求項 8】 請求項 1、請求項 2、又は請求項 3 に記載の記憶デバイス制御装置において、

前記フェイルオーバー時に引き継がれるデータには、少なくとも

N F S ユーザデータ、C I F S ユーザデータ、装置管理者データ、フェイルオーバー用ハートビート、チャンネル制御部の I P アドレス、N F S ファイルロック情報、クラスタ制御情報のいずれかが含まれることを特徴とする記憶デバイス制御装置。

【請求項 9】 情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応する I / O 要求を出力する I / O プロセッサとが形成された回路基板を有する複数のチャンネル制御部を備え、

前記各チャンネル制御部はフェイルオーバーを行うためのグループに類別されている記憶デバイス制御装置に、

前記各チャンネル制御部により更新される前記フェイルオーバー時に引き継がれるデータを、当該チャンネル制御部と同一の前記グループに類別されている前記各チャンネル制御部が共通にアクセス可能な前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域である共有ボリュームに記憶するステップを実行させるためのプログラム。

【請求項 1 0】 情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応する I / O 要求を出力する I / O プロセッサとが形成された回

路基板を有する複数のチャネル制御部を備え、

前記各チャネル制御部はフェイルオーバーを行うためのグループに類別されている記憶デバイス制御装置に、

前記各チャネル制御部により更新される前記フェイルオーバー時に引き継がれるデータを、前記各チャネル制御部が共通にアクセス可能な前記記憶デバイス制御装置が備える共有メモリに記憶するステップを実行させるためのプログラム。

【請求項 1 1】 情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応する I / O 要求を出力する I / O プロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、

前記各チャネル制御部はフェイルオーバーを行うためのグループに類別されている記憶デバイス制御装置に、

前記各チャネル制御部により更新される前記フェイルオーバー時に引き継がれるデータを、当該チャネル制御部と同一の前記グループに類別されている他の前記チャネル制御部に、前記各チャネル制御部を相互に接続するネットワークを介して送信するステップを実行させるためのプログラム。

【請求項 1 2】 請求項 9 又は請求項 1 0 に記載のプログラムにおいて、

前記各チャネル制御部には、当該チャネル制御部が個別にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域であるローカルボリュームがそれぞれ割り当てられており、

前記データを更新した前記チャネル制御部と同一の前記グループに類別されている他の前記チャネル制御部の前記ローカルボリュームにも前記データを記憶するステップをさらに実行させるためのプログラム。

【請求項 1 3】 請求項 9 又は請求項 1 0 に記載のプログラムにおいて、

前記各チャネル制御部には、当該チャネル制御部が個別にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域であるローカルボリュームがそれぞれ割り当てられており、

前記データを更新した前記チャネル制御部と同一の前記グループに類別されている他の前記チャネル制御部の前記ローカルボリュームにも前記データを記憶す

るステップと、

前記データの参照先が記録された引き継ぎデータ参照テーブルを参照するステップと、

前記引き継ぎデータ参照テーブルに記録された前記データの参照先に基づき、前記共有ボリューム、前記共有メモリ、又は前記ローカルボリュームのいずれかから前記データを読み出すステップと  
をさらに実行させるためのプログラム。

【請求項 14】 請求項 11 に記載のプログラムにおいて、

前記データが前記記憶デバイス制御装置内の全ての前記チャンネル制御部に共通に参照されるデータである場合には、当該データを、前記ネットワークを介して、前記記憶デバイス制御装置内の全ての前記チャンネル制御部に送信するステップを前記記憶デバイス制御装置に実行させるためのプログラム。

【請求項 15】 請求項 9 に記載のプログラムにおいて、

前記データが前記記憶デバイス制御装置内の全ての前記チャンネル制御部に共通に参照されるデータである場合には、当該データを、前記記憶デバイス制御装置内の全ての前記チャンネル制御部が共通にアクセス可能な、前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域である第 2 の共有ボリュームに記憶するステップを前記記憶デバイス制御装置に実行させるためのプログラム。

【請求項 16】 請求項 9、請求項 10、又は請求項 11 に記載のプログラムにおいて、

前記フェイルオーバー時に引き継がれるデータには、少なくとも

NFS ユーザデータ、CIFS ユーザデータ、装置管理者データ、フェイルオーバー用ハートビート、チャンネル制御部の IP アドレス、NFS ファイルロック情報、クラスタ制御情報のいずれかが含まれることを特徴とするプログラム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、記憶デバイス制御装置、及びプログラムに関する。



**【 0 0 0 2 】****【従来の技術】**

近年コンピュータシステムで取り扱われるデータ量が急激に増加している。このようなデータを管理するためのストレージシステムとして、最近ではミッドレンジクラスやエンタープライズクラスと呼ばれるような、巨大な記憶資源を提供する R A I D (Redundant Arrays of Inexpensive Disks) 方式で管理された大規模なストレージシステムが注目されている。

**【 0 0 0 3 】**

また、ストレージシステムと情報処理装置とを T C P / I P (Transmission Control Protocol/Internet Protocol) プロトコル等を用いたネットワークで相互に接続し、情報処理装置からのファイルレベルでのアクセスを実現する N A S (Network Attached Storage) と呼ばれるストレージシステムが開発されている。

**【 0 0 0 4 】**

一方、ストレージシステムが提供する記憶領域を使用する情報処理装置に障害が発生しても、他の情報処理装置に処理を代替させることにより、情報処理装置によるサービスの提供を継続させることができるフェイルオーバと呼ばれる技術が開発されている。

**【 0 0 0 5 】****【特許文献 1】**

特開平 8 - 2 1 2 0 9 5 号公報

**【 0 0 0 6 】****【発明が解決しようとする課題】**

しかしながら従来のストレージシステムにおけるフェイルオーバでは、情報処理装置間を跨る共通の記憶領域が設けられていないため、フェイルオーバを行う場合には、元の情報処理装置が使用していた記憶領域のデータを、処理を代替する情報処理装置が使用できるように引き継ぎ処理をしなければならない。また情報処理装置に障害が発生した後に記憶領域のデータの引き継ぎが行われるため、代替処理が開始されるまでにタイムラグが発生していた。さらにデータの引き継

ぎのための処理が煩雑であった。

本発明は上記課題を鑑みてなされたものであり、記憶デバイス制御装置、及びプログラムを提供することを主たる目的とする。

#### 【0007】

##### 【課題を解決するための手段】

上記課題を解決するために、本発明に係る記憶デバイス制御装置は、情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対する前記データ入出力要求に対応する I/O 要求を出力する I/O プロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、前記各チャネル制御部はフェイルオーバを行うためのグループに類別されている記憶デバイス制御装置であって、前記各チャネル制御部により更新される前記フェイルオーバ時に引き継がれるデータを、当該チャネル制御部と同一の前記グループに類別されている前記各チャネル制御部が共通にアクセス可能な前記記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域である共有ボリュームに記憶する手段を備える。

#### 【0008】

なお、前記情報処理装置とは、前記構成の前記記憶デバイス制御装置を備えて構成されるストレージシステムに LAN (Local Area Network) を介してアクセスする、例えば、パーソナルコンピュータやメインフレームコンピュータである。ファイルアクセス処理部の機能は CPU 上で実行されるオペレーティングシステムおよびこのオペレーティングシステム上で動作する例えば NFS (Network File System) 等のソフトウェアによって提供される。記憶デバイスは例えばハードディスク装置などのディスクドライブである。I/O プロセッサは例えばファイルアクセス処理部のハードウェア要素である前記 CPU とは独立した IC (Integrated Circuit) をハードウェア要素とし、ファイルアクセス処理部とディスク制御部との間の通信を制御する。ディスク制御部は、記憶デバイスに対してデータの書き込みや読み出しを行う。

その他、本願が開示する課題、及びその解決方法は、発明の実施の形態の欄、及び図面により明らかにされる。

**【 0 0 0 9 】****【発明の実施の形態】**

以下、本発明の実施の形態について図面を用いて詳細に説明する。

まず、本実施の形態に係るストレージシステム 6 0 0 の全体構成を示すブロック図を図 1 に示す。

===全体構成例===

ストレージシステム 6 0 0 は、記憶デバイス制御装置 1 0 0 と記憶デバイス 3 0 0 とを備えている。記憶デバイス制御装置 1 0 0 は、情報処理装置 2 0 0 から受信したコマンドに従って記憶デバイス 3 0 0 に対する制御を行う。例えば情報処理装置 2 0 0 からデータの入出力要求を受信して、記憶デバイス 3 0 0 に記憶されているデータの入出力のための処理を行う。データは、記憶デバイス 3 0 0 が備えるディスクドライブにより提供される物理的な記憶領域上に論理的に設定される記憶領域である論理ボリューム (Logical Unit) (以下、L U と記す) に記憶されている。また記憶デバイス制御装置 1 0 0 は、情報処理装置 2 0 0 との間で、ストレージシステム 6 0 0 を管理するための各種コマンドの授受も行う。

**【 0 0 1 0 】**

情報処理装置 2 0 0 は C P U (Central Processing Unit) やメモリを備えたコンピュータである。情報処理装置 2 0 0 が備える C P U により各種プログラムが実行されることにより様々な機能が実現される。情報処理装置 2 0 0 は、例えばパーソナルコンピュータやワークステーションであることもあるし、メインフレームコンピュータであることもある。

**【 0 0 1 1 】**

図 1 において、情報処理装置 2 0 0 は、L A N (Local Area Network) 4 0 0 を介して記憶デバイス制御装置 1 0 0 と接続されている。L A N 4 0 0 は、インターネットとすることもできるし、専用のネットワークとすることもできる。L A N 4 0 0 を介して行われる情報処理装置 2 0 0 と記憶デバイス制御装置 1 0 0 との間の通信は、例えば T C P / I P プロトコルに従って行われる。情報処理装置 2 0 0 からは、ストレージシステム 6 0 0 に対して、ファイル名指定によるデータアクセス要求 (ファイル単位でのデータ入出力要求。以下、ファイルアクセ

ス要求と記す) が送信される。

#### 【0012】

記憶デバイス制御装置100は、チャンネル制御部110を備える。以下、このチャンネル制御部110のことをCHNとも記す。記憶デバイス制御装置100は、チャンネル制御部110によりLAN400を介して情報処理装置200との間で通信を行う。チャンネル制御部110は、情報処理装置200からのファイルアクセス要求を個々に受け付ける。すなわち、チャンネル制御部110には、個々にLAN400上のネットワークアドレス（例えば、IPアドレス）が割り当てられていてそれぞれが個別にNASとして振る舞い、個々のNASがあたかも独立したNASが存在しているかのようにNASのサービスを情報処理装置200に提供することができる。以下、チャンネル制御部110をCHNと記す。このように1台のストレージシステム600に個別にNASとしてのサービスを提供するチャンネル制御部110を備えるように構成したことで、従来、独立したコンピュータで個別に運用されていたNASサーバが一台のストレージシステム600に集約されて運用される。そして、この構成によってストレージシステム600の統括的な管理が可能となり、各種設定・制御や障害管理、バージョン管理といった保守業務の効率化が図られる。

#### 【0013】

なお、本実施の形態に係る記憶デバイス制御装置100のチャンネル制御部110の機能は、後述するように、一体的にユニット化された回路基板上に形成されたハードウェア及びこのハードウェアにより実行されるオペレーティングシステム（以下、OSと記す）やこのOS上で動作するアプリケーションプログラムなどのソフトウェアにより実現される。このように本実施例のストレージシステム600では、従来ハードウェアの一部として実装されてきた機能が主としてソフトウェアによって実現されている。このため、本実施例のストレージシステム600では柔軟性に富んだシステム運用が可能となり、多様で変化の激しいユーザーニーズに対応したきめ細かなサービスを提供することが可能となる。

#### 【0014】

===記憶デバイス===

記憶デバイス 300 は多数のディスクドライブ（物理ディスク）を備えており、情報処理装置 200 に対して記憶領域を提供する。データは、ディスクドライブにより提供される物理的な記憶領域上に論理的に設定される記憶領域である LU に記憶されている。ディスクドライブとしては、例えばハードディスク装置やフレキシブルディスク装置、半導体記憶装置等様々なものを用いることができる。

#### 【0015】

なお、記憶デバイス 300 は例えば複数のディスクドライブによりディスクアレイを構成するようにすることもできる。この場合、情報処理装置 200 に対して提供される記憶領域は、RAID により管理された複数のディスクドライブにより提供されるようにすることもできる。

記憶デバイス制御装置 100 と記憶デバイス 300 との間は図 1 のように直接に接続される形態とすることもできるし、ネットワークを介して接続するようにすることもできる。さらに記憶デバイス 300 は記憶デバイス制御装置 100 と一体として構成されることもできる。

#### 【0016】

記憶デバイス 300 に設定される LU には、情報処理装置 200 からアクセス可能なユーザ LU や、チャンネル制御部 110 の制御のために使用されるシステム LU 等がある。システム LU には CHN 110 で実行されるオペレーティングシステムも格納される。また各 LU にはチャンネル制御部 110 が対応付けられている。これによりチャンネル制御部 110 毎にアクセス可能な LU が割り当てられている。また上記対応付けは、複数のチャンネル制御部 110 で一つの LU を共有するようにすることもできる。なお以下において、ユーザ LU やシステム LU をユーザディスク、システムディスク等とも記す。また、複数のチャンネル制御部 110 で共有される LU を共有 LU あるいは共有ディスクと記す。

#### 【0017】

=== 記憶デバイス制御装置 ===

記憶デバイス制御装置 100 はチャンネル制御部 110、共有メモリ 120、キャッシュメモリ 130、ディスク制御部 140、管理端末 160、接続部 150 を備える。

**【0018】**

チャンネル制御部110は情報処理装置200との間で通信を行うための通信インタフェースを備え、情報処理装置200との間でデータ入出力コマンド等を授受する機能を備える。例えばCHN110は情報処理装置1乃至3(200)からのファイルアクセス要求を受け付ける。そしてファイルの記憶アドレスやデータ長等を求めて、ファイルアクセス要求に対応するI/O要求を出力することにより、記憶デバイス300へのアクセスを行う。これによりストレージシステム600はNASとしてのサービスを情報処理装置1乃至3(200)に提供することができる。なおI/O要求にはデータの先頭アドレス、データ長、読み出し又は書き込み等のアクセスの種別が含まれている。またデータの書き込みの場合にはI/O要求には書き込みデータが含まれているようにすることもできる。I/O要求の出力は、後述するI/Oプロセッサ119により行われる。

**【0019】**

各チャンネル制御部110は管理端末160と共に内部LAN151で接続されている。これによりチャンネル制御部110に実行させるマイクロプログラム等を管理端末160から送信しインストールすることが可能となっている。チャンネル制御部110の構成については後述する。

**【0020】**

接続部150はチャンネル制御部110、共有メモリ120、キャッシュメモリ130、ディスク制御部140を相互に接続する。チャンネル制御部110、共有メモリ120、キャッシュメモリ130、ディスク制御部140間でのデータやコマンドの授受は接続部150を介することにより行われる。接続部150は例えば高速スイッチングによりデータ伝送を行う超高速クロスバススイッチなどの高速バスである。チャンネル制御部110同士が高速バスで接続されていることで、個々のコンピュータ上で動作するNASサーバがLANを通じて接続する従来の構成に比べてチャンネル制御部110間の通信パフォーマンスが大幅に向上する。またこれにより高速なファイル共有機能や高速フェイルオーバーなども可能となる。

**【0021】**

共有メモリ 120 及びキャッシュメモリ 130 は、チャンネル制御部 110、ディスク制御部 140 により共有される記憶メモリである。共有メモリ 120 は主に制御情報やコマンド等を記憶するために利用されるのに対し、キャッシュメモリ 130 は主にデータを記憶するために利用される。

#### 【0022】

例えば、あるチャンネル制御部 110 が情報処理装置 200 から受信したデータ入出力コマンドが書き込みコマンドであった場合には、当該チャンネル制御部 110 は書き込みコマンドを共有メモリ 120 に書き込むと共に、情報処理装置 200 から受信した書き込みデータをキャッシュメモリ 130 に書き込む。一方、ディスク制御部 140 は共有メモリ 120 を監視しており、共有メモリ 120 に書き込みコマンドが書き込まれたことを検出すると、当該コマンドに従ってキャッシュメモリ 130 から書き込みデータを読み出して記憶デバイス 300 に書き込む。

#### 【0023】

またあるチャンネル制御部 110 が情報処理装置 200 から受信したデータ入出力コマンドが読み出しコマンドであった場合には、当該チャンネル制御部 110 は読み出しコマンドを共有メモリ 120 に書き込むと共に、読み出し対象となるデータがキャッシュメモリ 130 に存在するかどうかを調べる。ここでキャッシュメモリ 130 に存在すれば、チャンネル制御部 110 はそのデータを情報処理装置 200 に送信する。一方、読みだし対象となるデータがキャッシュメモリ 130 に存在しない場合には、共有メモリ 120 を監視することにより読み出しコマンドが共有メモリ 120 に書き込まれたことを検出したディスク制御部 140 が、記憶デバイス 300 から読みだし対象となるデータを読み出してこれをキャッシュメモリ 130 に書き込むと共に、その旨を共有メモリ 120 に書き込む。そして、チャンネル制御部 110 は共有メモリ 120 を監視することにより読みだし対象となるデータがキャッシュメモリ 130 に書き込まれたことを検出すると、そのデータを情報処理装置 200 に送信する。

#### 【0024】

なお、このようにチャンネル制御部 110 からディスク制御部 140 に対するデ

ータの書き込みや読み出しの指示を共有メモリ 120 を介在させて間接に行う構成の他、例えばチャネル制御部 110 からディスク制御部 140 に対してデータの書き込みや読み出しの指示を共有メモリ 120 を介さずに直接に行う構成とすることもできる。

#### 【0025】

ディスク制御部 140 は記憶デバイス 300 の制御を行う。例えば上述のように、チャネル制御部 110 が情報処理装置 200 から受信したデータ書き込みコマンドに従って記憶デバイス 300 へデータの書き込みを行う。また、チャネル制御部 110 により送信された論理アドレス指定による LU へのデータアクセス要求を、物理アドレス指定による物理ディスクへのデータアクセス要求に変換する。記憶デバイス 300 における物理ディスクが RAID により管理されている場合には、RAID 構成（例えば、RAID 0, 1, 5）に従ったデータのアクセスを行う。またディスク制御部 140 は、記憶デバイス 300 に記憶されたデータの複製管理の制御やバックアップ制御を行う。さらにディスク制御部 140 は、災害発生時のデータ消失防止（ディザスタリカバリ）などを目的としてプライマリサイトのストレージシステム 600 のデータの複製をセカンダリサイトに設置された他のストレージシステムにも記憶する制御（レプリケーション機能、又はリモートコピー機能）なども行う。

#### 【0026】

各ディスク制御部 140 は管理端末 160 と共に内部 LAN 151 で接続されており、相互に通信を行うことが可能である。これにより、ディスク制御部 140 に実行させるマイクロプログラム等を管理端末 160 から送信しインストールすることが可能となっている。ディスク制御部 140 の構成については後述する。

#### 【0027】

本実施例においては、共有メモリ 120 及びキャッシュメモリ 130 がチャネル制御部 110 及びディスク制御部 140 に対して独立に設けられていることについて記載したが、本実施例はこの場合に限られるものでなく、共有メモリ 120 又はキャッシュメモリ 130 がチャネル制御部 110 及びディスク制御部 14



0の各々に分散されて設けられることも好ましい。この場合、接続部150は、分散された共有メモリ又はキャッシュメモリを有するチャンネル制御部110及びディスク制御部140を相互に接続させることになる。

#### 【0028】

===管理端末===

管理端末160はストレージシステム600を保守・管理するためのコンピュータである。管理端末160を操作することにより、例えば記憶デバイス300内の物理ディスク構成の設定や、LUの設定、チャンネル制御部110において実行されるマイクロプログラムのインストール等を行うことができる。ここで、記憶デバイス300内の物理ディスク構成の設定としては、例えば物理ディスクの増設や減設、RAID構成の変更（例えばRAID1からRAID5への変更等）等を行うことができる。さらに管理端末160からは、ストレージシステム600の動作状態の確認や故障部位の特定、チャンネル制御部110で実行されるオペレーティングシステムのインストール等の作業を行うこともできる。また管理端末160はLANや電話回線等で外部保守センタと接続されており、管理端末160を利用してストレージシステム600の障害監視を行ったり、障害が発生した場合に迅速に対応することも可能である。障害の発生は例えばOSやアプリケーションプログラム、ドライバソフトウェアなどから通知される。この通知はHTTPプロトコルやSNMP（Simple Network Management Protocol）、電子メールなどにより行われる。これらの設定や制御は、管理端末160で動作するWebサーバが提供するWebページをユーザインタフェースとしてオペレータなどにより行われる。オペレータ等は、管理端末160を操作して障害監視する対象や内容の設定、障害通知先の設定などを行うこともできる。

#### 【0029】

管理端末160は記憶デバイス制御装置100に内蔵されている形態とすることもできるし、外付けされている形態とすることもできる。また管理端末160は、記憶デバイス制御装置100及び記憶デバイス300の保守・管理を専用に行うコンピュータとすることもできるし、汎用のコンピュータに保守・管理機能を持たせたものとすることもできる。

**【0030】**

管理端末160の構成を示すブロック図を図2に示す。

管理端末160は、CPU161、メモリ162、ポート163、記録媒体読取装置164、入力装置165、出力装置166、記憶装置168を備える。

**【0031】**

CPU161は管理端末160の全体の制御を司るもので、メモリ162に格納されたプログラム162cを実行することにより上記Webサーバとしての機能等を実現する。メモリ162には、物理ディスク管理テーブル162aとLU管理テーブル162bとプログラム162cとが記憶されている。

**【0032】**

物理ディスク管理テーブル162aは、記憶デバイス300に備えられる物理ディスク（ディスクドライブ）を管理するためのテーブルである。物理ディスク管理テーブル162aを図3に示す。図3においては、記憶デバイス300が備える多数の物理ディスクのうち、ディスク番号#001乃至#006までが示されている。それぞれの物理ディスクに対して、容量、RAID構成、使用状況が示されている。

**【0033】**

LU管理テーブル162bは、上記物理ディスク上に論理的に設定されるLUを管理するためのテーブルである。LU管理テーブル162bを図4に示す。図4においては、記憶デバイス300上に設定される多数のLUのうち、LU番号#1乃至#3までが示されている。それぞれのLUに対して、物理ディスク番号、容量、RAID構成が示されている。

**【0034】**

記録媒体読取装置164は、記録媒体167に記録されているプログラムやデータを読み取るための装置である。読み取られたプログラムやデータはメモリ162や記憶装置168に格納される。従って、例えば記録媒体167に記録されたプログラム162cを、記録媒体読取装置164を用いて上記記録媒体167から読み取って、メモリ162や記憶装置168に格納することができる。記録媒体167としてはフレキシブルディスクやCD-ROM、DVD-

ROM、DVD-RAM、半導体メモリ等を用いることができる。なお、上記プログラム162cは管理端末160を動作させるためのプログラムとすることができる他、チャンネル制御部110やディスク制御部140にOS701やアプリケーションプログラムをインストールするためのプログラムや、バージョンアップするためのプログラムとすることもできる。記録媒体読取装置164は管理端末160に内蔵されている形態とすることもできるし、外付されている形態とすることもできる。記憶装置168は、例えばハードディスク装置やフレキシブルディスク装置、半導体記憶装置等である。入力装置165はオペレータ等による管理端末160へのデータ入力等のために用いられる。入力装置165としては例えばキーボードやマウス等が用いられる。出力装置166は情報を外部に出力するための装置である。出力装置166としては例えばディスプレイやプリンタ等が用いられる。ポート163は内部LAN151に接続されており、これにより管理端末160はチャンネル制御部110やディスク制御部140等と通信を行うことができる。またポート163は、LAN400に接続するようにすることもできるし、電話回線に接続するようにすることもできる。

#### 【0035】

===外観図===

次に、本実施の形態に係るストレージシステム600の外観構成を図5に示す。また、記憶デバイス制御装置100の外観構成を図6に示す。

図5に示すように、本実施の形態に係るストレージシステム600は記憶デバイス制御装置100及び記憶デバイス300がそれぞれの筐体に納められた形態をしている。記憶デバイス制御装置100の筐体の両側に記憶デバイス300の筐体が配置されている。

#### 【0036】

記憶デバイス制御装置100は、正面中央部に管理端末160が備えられている。管理端末160はカバーで覆われており、図6に示すようにカバーを開けることにより管理端末160を使用することができる。なお図6に示した管理端末160はいわゆるノート型パーソナルコンピュータの形態をしているが、どのような形態とすることも可能である。

## 【0037】

管理端末160の下部には、チャンネル制御部110のボードを装着するためのスロットが設けられている。チャンネル制御部110のボードとは、チャンネル制御部110の回路基板が形成されたユニットであり、スロットへの装着単位である。本実施の形態に係るストレージシステム600においては、スロットは8つあり、図5及び図6には8つのスロットにチャンネル制御部110のボードが装着された状態が示されている。各スロットにはチャンネル制御部110のボードを装着するためのガイドレールが設けられている。ガイドレールに沿ってチャンネル制御部110のボードをスロットに挿入することにより、チャンネル制御部110のボードを記憶デバイス制御装置100に装着することができる。また各スロットに装着されたチャンネル制御部110のボードは、ガイドレールに沿って手前方向に引き抜くことにより取り外すことができる。また各スロットの奥手方向正面部には、各チャンネル制御部110のボードを記憶デバイス制御装置100と電氣的に接続するためのコネクタが設けられている。

## 【0038】

なお、スロットには以上に説明したNASとして機能するタイプのチャンネル制御部110以外にも、SAN (Storage Area Network) に接続する機能を備えるタイプや、FICON (Fibre Connection) (登録商標) やESCON (Enterprise System Connection) (登録商標) 等のメインフレーム系のプロトコルに従って通信を行う機能を備えるタイプのチャンネル制御部110が装着されることもある。またチャンネル制御部110のボードを装着しないスロットを設けるようにすることもできる。

## 【0039】

各スロットのチャンネル制御部110は、同種の複数のチャンネル制御部110でクラスタを構成する。例えば2枚のCHN110をペアとしてクラスタを構成することができる。クラスタを構成することにより、クラスタ内のあるチャンネル制御部110に障害が発生した場合でも、障害が発生したチャンネル制御部110がそれまで行っていた処理をクラスタ内の他のチャンネル制御部110に引き継ぐようにすることができる (フェイルオーバー制御)。2枚のCHN110でクラスタ

を構成している様子を示す図を図 11 に示すが、詳細は後述する。

#### 【0040】

なお、記憶デバイス制御装置 100 は信頼性向上のため電源供給が 2 系統化されており、チャンネル制御部 110 のボードが装着される上記 8 つのスロットは電源系統毎に 4 つずつに分けられている。そこでクラスタを構成する場合には、両方の電源系統のチャンネル制御部 110 のボードを含むようにする。これにより、片方の電源系統に障害が発生し電力の供給が停止しても、同一クラスタを構成する他方の電源系統に属するチャンネル制御部 110 のボードへの電源供給は継続されるため、当該チャンネル制御部 110 に処理を引き継ぐ（フェイルオーバー）ことができる。

#### 【0041】

なお、上述したように、チャンネル制御部 110 は上記各スロットに装着可能なボードとして提供されるが、上記一つのボードは一体形成された複数枚数の回路基板から構成されているようにすることもできる。

#### 【0042】

ディスク制御部 140 や共有メモリ 120 等の、記憶デバイス制御装置 100 を構成する他の装置については図 5 及び図 6 には示されていないが、記憶デバイス制御装置 100 の背面側等に装着されている。

また記憶デバイス制御装置 100 には、チャンネル制御部 110 のボード等から発生する熱を放出するためのファン 170 が設けられている。ファン 170 は記憶デバイス制御装置 100 の上面部に設けられる他、チャンネル制御部 110 用スロットの上部にも設けられている。

#### 【0043】

ところで、筐体に収容されて構成される記憶デバイス制御装置 100 および記憶デバイス 300 としては、例えば SAN 対応として製品化されている従来構成の装置を利用することができる。特に上記のように CHN 110 のボードのコネクタ形状を従来構成の筐体に設けられているスロットにそのまま装着できる形状とすることで従来構成の装置をより簡単に利用することができる。つまり本実施例のストレージシステム 600 は、既存の製品を利用することで容易に構築する

ことができる。

#### 【0044】

===チャンネル制御部===

本実施の形態に係るストレージシステム600は、上述の通りCHN110により情報処理装置200からのファイルアクセス要求を受け付け、NASとしてのサービスを情報処理装置200に提供する。

#### 【0045】

CHN110のハードウェア構成を図7に示す。この図に示すようにCHN110のハードウェアは一体的にユニット化されたボードで構成される。以下、このユニットのことをNASボードとも記す。NASボードは一枚もしくは複数枚の回路基板を含んで構成される。より具体的には、NASボードは、ネットワークインタフェース部111、CPU112、メモリ113、入出力制御部114、I/O (Input/Output) プロセッサ119、NVRAM (Non Volatile RAM) 115、ボード接続用コネクタ116、通信コネクタ117を備え、これらが同一のユニット化された回路基板として形成されて構成されている。

#### 【0046】

ネットワークインタフェース部111は、情報処理装置200との間で通信を行うための通信インタフェースを備えている。CHN110の場合は、例えばTCP/IPプロトコルに従って情報処理装置200から送信されたファイルアクセス要求を受信する。通信コネクタ117は情報処理装置200と通信を行うためのコネクタである。CHN110の場合はLAN400に接続可能なコネクタであり、例えばイーサネット（登録商標）に対応している。

#### 【0047】

CPU112は、CHN110をNASボードとして機能させるための制御を司る。

メモリ113には様々なプログラムやデータが記憶される。例えば図8に示すメタデータ730やロックテーブル720、また図10に示すNASマネージャ706等の各種プログラムが記憶される。

#### 【0048】

メタデータ 730 はファイルシステムプログラム 703 により実現されるファイルシステムが管理しているファイルに対応させて生成される情報である。メタデータ 730 には例えばファイルのデータが記憶されている LU 上のアドレスやデータサイズなど、ファイルの保管場所を特定するための情報が含まれる。メタデータ 730 にはファイルの容量、所有者、更新時刻等の情報が含まれることもある。また、メタデータ 730 はファイルだけでなくディレクトリに対応させて生成されることもある。メタデータ 730 の例を図 12 に示す。メタデータ 730 は記憶デバイス 300 上の各 LU にも記憶されている。

#### 【0049】

ロックテーブル 720 は、情報処理装置 200 からのファイルアクセスに対して排他制御を行うためのテーブルである。排他制御を行うことにより情報処理装置 200 でファイルを共用することができる。ロックテーブル 720 を図 13 に示す。図 13 に示すようにロックテーブル 720 にはファイルロックテーブル 721 と LU ロックテーブル 722 とがある。ファイルロックテーブル 721 は、ファイル毎にロックが掛けられているか否かを示すためのテーブルである。いずれかの情報処理装置 200 によりあるファイルがオープンされている場合に当該ファイルにロックが掛けられる。ロックが掛けられたファイルに対する他の情報処理装置 200 によるアクセスは禁止される。LU ロックテーブル 722 は、LU 毎にロックが掛けられているか否かを示すためのテーブルである。いずれかの情報処理装置 200 により、ある LU に対するアクセスが行われている場合に当該 LU にロックが掛けられる。ロックが掛けられた LU に対する他の情報処理装置 200 によるアクセスは禁止される。

#### 【0050】

入出力制御部 114 は、ディスク制御部 140 やキャッシュメモリ 130、共有メモリ 120、管理端末 160 との間でデータやコマンドの授受を行う。入出力制御部 114 は I/O プロセッサ 119 や NVRAM 115 を備えている。I/O プロセッサ 119 は例えば 1 チップのマイコンで構成される。I/O プロセッサ 119 は上記データやコマンドの授受を制御し、CPU 112 とディスク制御部 140 との間の通信を中継する。NVRAM 115 は I/O プロセッサ 11

9の制御を司るプログラムを格納する不揮発性メモリである。NVRAM115に記憶されるプログラムの内容は、管理端末160や、後述するNASマネージャ706からの指示により書き込みや書き換えを行うことができる。

#### 【0051】

次にディスク制御部140のハードウェア構成を示す図を図9に示す。

ディスク制御部140は、一体的にユニット化されたボードとして形成されている。ディスク制御部140のボードは、インタフェース部141、メモリ143、CPU142、NVRAM144、ボード接続用コネクタ145を備え、これらが一体的にユニット化された回路基板として形成されている。

#### 【0052】

インタフェース部141は、接続部150を介してチャネル制御部110等との間で通信を行うための通信インタフェースや、記憶デバイス300との間で通信を行うための通信インタフェースを備えている。

CPU142は、ディスク制御部140全体の制御を司ると共に、チャネル制御部140や記憶デバイス300、管理端末160との間の通信を行う。メモリ143やNVRAM144に格納された各種プログラムを実行することにより本実施の形態に係るディスク制御部140の機能が実現される。ディスク制御部140により実現される機能としては、記憶デバイス300の制御やRAID制御、記憶デバイス300に記憶されたデータの複製管理やバックアップ制御、リモートコピー制御等である。

#### 【0053】

NVRAM144はCPU142の制御を司るプログラムを格納する不揮発性メモリである。NVRAM144に記憶されるプログラムの内容は、管理端末160や、NASマネージャ706からの指示により書き込みや書き換えを行うことができる。

またディスク制御部140のボードはボード接続用コネクタ145を備えている。ボード接続用コネクタ145が記憶デバイス制御装置100側のコネクタと嵌合することにより、ディスク制御部140のボードは記憶デバイス制御装置100と電氣的に接続される。



**【0054】**

===ソフトウェア構成===

次に、本実施の形態に係るストレージシステム600におけるソフトウェア構成図を図10に示す。

オペレーティングシステム701は例えばUNIX（登録商標）である。オペレーティングシステム701上では、RAIDマネージャ708、ボリュームマネージャ707、SVPマネージャ709、ファイルシステムプログラム703、ネットワーク制御部702、障害管理プログラム705、NASマネージャ706などのソフトウェアが動作する。

**【0055】**

オペレーティングシステム701上で動作するRAIDマネージャ708は、RAID制御部740に対するパラメータの設定やRAID制御部740を制御する機能を提供する。RAIDマネージャ708はオペレーティングシステム701やオペレーティングシステム701上で動作する他のアプリケーション、もしくは管理端末160からパラメータや制御指示情報を受け付けて、受け付けたパラメータのRAID制御部740への設定や、RAID制御部指示情報に対応する制御コマンドの送信を行う。

**【0056】**

ここで設定されるパラメータとしては、例えば、RAIDグループを構成する記憶デバイス（物理ディスク）を定義（RAIDグループの構成情報、ストライプサイズの指定など）するためのパラメータ、RAIDレベル（例えば0, 1, 5）を設定するためのパラメータなどがある。また、RAIDマネージャ708がRAID制御部740に送信する制御コマンドとしてはRAIDの構成・削除・容量変更を指示するコマンド、各RAIDグループの構成情報を要求するコマンドなどがある。

**【0057】**

ボリュームマネージャ707は、RAID制御部740によって提供されるLUをさらに仮想化した仮想化論理ボリュームをファイルシステムプログラム703に提供する。1つの仮想化論理ボリュームは1以上の論理ボリュームによって

構成される。

#### 【 0 0 5 8 】

ファイルシステムプログラム 7 0 3 の主な機能は、ネットワーク制御部 7 0 2 が受信したファイルアクセス要求に指定されているファイル名とそのファイル名が格納されている仮想化論理ボリューム上のアドレスとの対応づけを管理することである。例えば、ファイルシステムプログラム 7 0 3 はファイルアクセス要求に指定されているファイル名に対応する仮想化論理ボリューム上のアドレスを特定する。

#### 【 0 0 5 9 】

ネットワーク制御部 7 0 2 は、N F S (Network File System) 7 1 1 と C I F S (Common Interface File System) 7 1 3 の 2 つのファイルシステムプロトコルを含んで構成される。N F S 7 1 1 は、N F S 7 1 1 が動作する U N I X (登録商標) 系の情報処理装置 2 0 0 からのファイルアクセス要求を受け付ける。一方、C I F S 7 1 3 は C I F S 7 1 3 が動作する W i n d o w s (登録商標) 系の情報処理装置 2 0 0 からのファイルアクセス要求を受け付ける。

#### 【 0 0 6 0 】

N A S マネージャ 7 0 6 は、ストレージシステム 6 0 0 について、その動作状態の確認、設定や制御などを行うためのプログラムである。N A S マネージャ 7 0 6 は W e b サーバとしての機能も有し、情報処理装置 2 0 0 からストレージシステム 6 0 0 の設定や制御を行うための設定 W e b ページを情報処理装置 2 0 0 に提供する。設定 W e b ページはチャネル制御部 1 1 0 の個々において動作する N A S マネージャ 7 0 6 により提供される。N A S マネージャ 7 0 6 は、情報処理装置 2 0 0 からの H T T P (HyperText Transport Protocol) リクエストに応じて、設定 W e b ページのデータを情報処理装置 2 0 0 に送信する。情報処理装置 2 0 0 に表示された設定 W e b ページを利用してシステムアドミニストレータなどによりストレージシステム 6 0 0 の設定や制御の指示が行われる。

#### 【 0 0 6 1 】

N A S マネージャ 7 0 6 は、設定 W e b ページに対する操作に起因して情報処理装置 2 0 0 から送信される設定や制御に関するデータを受信してそのデータに

対応する設定や制御を実行する。これにより、情報処理装置 2 0 0 からストレージシステム 6 0 0 の様々な設定や制御を行うことができる。また N A S マネージャ 7 0 6 は設定 W e b ページの設定内容をチャンネル制御部 1 1 0 上で動作する O S やアプリケーションプログラム、ディスク制御部 1 4 0 等に通知する。設定 W e b ページで設定された内容は共有 L U 3 1 0 に管理されることもある。

#### 【 0 0 6 2 】

N A S マネージャ 7 0 6 の設定 W e b ページを利用して行うことができる内容としては、例えば、L U の管理や設定（容量管理や容量拡張・縮小、ユーザ割り当て等）、上述の複製管理やリモートコピー（レプリケーション）等の機能に関する設定や制御（複製元の L U と複製先の L U の設定など）、冗長構成された C H N でのクラスタの管理（フェイルオーバーさせる相手の対応関係の設定、フェイルオーバー方法など）、O S や O S 上で動作するアプリケーションプログラムのバージョン管理などがある。

#### 【 0 0 6 3 】

なお、N A S マネージャ 7 0 6 によるストレージシステム 6 0 0 の動作状態の確認、設定や制御には、上述した設定 W e b ページを介する方法以外にクライアント・サーバシステムとすることもできる。その場合、N A S マネージャ 7 0 6 はクライアント・サーバシステムのサーバ機能を有し、情報処理装置 2 0 0 のクライアント機能の操作に起因して送信される設定や制御を、上述した設定 W e b ページと同様に実施することで、ストレージシステム 6 0 0 の動作状態の確認、設定や制御を行う。

#### 【 0 0 6 4 】

S V P マネージャ 7 0 9 は、管理端末 1 6 0 からの要求に応じて各種のサービスを管理端末 1 6 0 に提供する。例えば、L U の設定内容や R A I D の設定内容等のストレージシステム 6 0 0 に関する各種設定内容の管理端末 1 6 0 への提供や、管理端末 1 6 0 から入力されたストレージシステム 6 0 0 に関する各種設定の反映等を行う。

#### 【 0 0 6 5 】

=== クラスタ間同期制御 ===

障害管理プログラム 705 は、クラスタを構成するチャンネル制御部 110 間でのフェイルオーバー制御を行うためのプログラムである。

2 枚の CHN 110 でクラスタ 180 が構成されている様子を示す図を図 11 に示す。図 11 では、CHN 1 (チャンネル制御部 1) 110 と CHN 2 (チャンネル制御部 2) 110 とでクラスタ (グループ) 180 が構成されている場合を示す。

上述したように、フェイルオーバー処理はクラスタ 180 を構成するチャンネル制御部 110 間で行われる。例えば CHN 1 (110) に何らかの障害が発生し、処理を継続することができなくなった場合には、CHN 1 (110) がそれまで行っていた処理は CHN 2 (110) に引き継がれ、CHN 2 (110) がその後の処理を行う。

#### 【0066】

なお、フェイルオーバーは CHN 110 に障害が発生したときに自動的に行われることもあるが、オペレータが管理端末 160 を操作して手動で行われることもある。またユーザが NAS マネージャ 706 により提供される設定 Web ページを利用することにより、情報処理装置 200 から手動で行われることもある。フェイルオーバーを手動で行う目的としては、耐用年数の経過やバージョンアップ、定期診断などのためにチャンネル制御部 110 のハードウェア (例えば NAS ボード) を交換する必要性が生じた場合などがある。

#### 【0067】

CHN 2 (110) が CHN 1 (110) の処理を引き継いで実行するためには、CHN 2 (110) は CHN 1 (110) から種々のデータを引き継ぐことが必要である。CHN 1 (110) から CHN 2 (110) に引き継がれるデータは、例えば NFS ユーザデータ、CIFS ユーザデータ、装置管理者データ、フェイルオーバー用ハートビート、CHN 1 (110) の IP アドレス、NFS ファイルロック情報、クラスタ制御情報等である。

#### 【0068】

NFS ユーザデータは、UNIX (登録商標) 系オペレーティングシステムが実行される情報処理装置 200 を利用して CHN 1 (110) によりファイルア

クセスサービスの提供を受けることができるユーザを管理するためのデータである。例えばユーザのログインIDやパスワード等が登録されるデータである。CHN2(110)はCHN1(110)のNFSユーザデータを引き継ぐことにより、それまでCHN1(110)によりファイルアクセスサービスの提供を受けていたユーザに対して、継続してファイルアクセスサービスを提供することが可能となる。

#### 【0069】

CIFSユーザデータは、Windows(登録商標)系オペレーティングシステムが実行される情報処理装置200を利用してCHN1(110)によりファイルアクセスサービスの提供を受けることができるユーザを管理するためのデータである。例えばユーザのログインIDやパスワード等が登録されるデータである。CHN2(110)はCHN1(110)のCIFSユーザデータを引き継ぐことにより、それまでCHN1(110)によりファイルアクセスサービスの提供を受けていたユーザに対して、継続してファイルアクセスサービスを提供することが可能となる。

#### 【0070】

装置管理者データは、ストレージシステム600あるいは記憶デバイス制御装置100の管理者を管理するためのデータである。例えば管理者のログインIDやパスワード、ホームディレクトリの位置を示すデータを含む。装置管理者データは、クラスタ180に関係なく、記憶デバイス制御装置100内の全チャネル制御部110に共通のデータである。

#### 【0071】

フェイルオーバー用ハートビートは、クラスタ180内の各CHN110が相互に動作状態を確認し合うためのデータである。CHN1(110)及びCHN2(110)は、相互に自己の処理が正常に行われていることを示すデータ(ハートビートマーク)を定期的に通知する。そして、相手側の通知有無を確認する。相手側による通知が確認できない場合には、相手側に何らかの障害が発生したと判断する。ハートビートマークには、CHN110の識別子や、CHN110が正常であるか異常であるかを示す符号、更新時刻等の情報が含まれる。

## 【0072】

CHN110のIPアドレスは、LAN400上でTCP/IP通信プロトコルに従って通信を行う場合にCHN110を識別するためのアドレスである。例えばCHN1(110)のIPアドレスをCHN2(110)が引き継ぐことにより、それまでCHN1(110)がLAN400を通じて受信していたデータをCHN2(110)が受信することができるようになる。

## 【0073】

NFSファイルロック情報は、ファイルロックテーブル721やLUロックテーブル722を含むファイルのロック状態を管理するためのデータである。

クラスタ制御情報は、クラスタ内のCHN110間で引き継ぎが必要な上記以外のデータを含む。例えば、障害が発生したCHN110が管理していたLUに構築されているファイルシステムのマウントに関するマウント情報やネットワークインタフェース部111のMAC(Media Access Control)アドレスやネットワークファイルシステムのエクスポート情報である。

CHN2(110)は、これらの引き継ぎデータをCHN1(110)から引き継ぐことにより、CHN1(110)がそれまで行っていた処理を引き継いで行う。

## 【0074】

本実施の形態における記憶デバイス制御装置100では、引き継ぎ元のCHN110と引き継ぎ先のCHN110との間で、これらの引き継ぎデータの同期を取るにより引き継ぎが行われる。すなわち、引き継ぎ元のCHN110と引き継ぎ先のCHN110の間では、引き継ぎデータの内容が同一となるように制御される。

例えば、あるCHN110により上記いずれかの引き継ぎデータが更新された場合には、更新された引き継ぎデータを、CHN110間を相互に接続するネットワークを介して他のCHN110に送信する。このようにすることにより、引き継ぎ元のCHN110と引き継ぎ先のCHN110とで参照される引き継ぎデータの内容を同一にすることができる。ここで、ネットワークとしてはLAN400を用いるようにすることもできるし、接続部150を用いるようにすること

もできる。また内部 LAN 151 を用いるようにすることもできる。

#### 【0075】

また複数の CHN 110 から共通にアクセス可能な共有 LU（共有ボリューム）に引き継ぎデータを記憶するようにすることにより引き継ぎデータの同期を実現することもできる。これにより、引き継ぎ元の CHN 110 と引き継ぎ先の CHN 110 とで同一の引き継ぎデータが参照されるようにできる。

また、複数の CHN 110 から共通にアクセス可能な共有メモリ 120 に引き継ぎデータを記憶するようにすることにより、引き継ぎ元の CHN 110 と引き継ぎ先の CHN 110 とで同一の引き継ぎデータが参照されるようにすることもできる。

#### 【0076】

本実施の形態における記憶デバイス制御装置 100 で行われる、引き継ぎデータの同期を説明するためのシステム構成図を図 14 及び図 15 に示す。

図 14 及び図 15 に示す記憶デバイス制御装置 100 では、CHN 1（110）と CHN 2（110）とでクラスタ A（180）が構成され、CHN 3（110）と CHN 4（110）とでクラスタ B（180）が構成されている。つまり、CHN 1 乃至 4（110）はクラスタ A（180）とクラスタ B（180）とに類別されている。そして各 CHN 110 は LAN 400 を介して情報処理装置 200 と接続されると共に、CHN 110 間も相互に接続される。また各 CHN 110 は接続部 150 を介して共有メモリ 120、システム LU、ユーザ LU、管理端末 160 と接続されている。図 14 においては、システム LU は CHN 110 毎に割り当てられていることが示されている。従って図 14 に示すシステム LU はローカル LU でもある。なお、図 14 及び図 15 に示す NAS 制御ソフトウェアは、障害管理プログラム 705 を含む。

#### 【0077】

図 15 は、ストレージシステム 600 が備える各 LU を、ローカル LU（ローカルボリューム）、共有 LU（共有ボリューム）、全体共有 LU（第 2 の共有ボリューム）に分類して記載したものである。ローカル LU は各 CHN 110 が個別にアクセス可能な LU である。共有 LU はクラスタ内の複数の CHN 110 か

ら共通にアクセス可能なLUである。全体共有LUは、ストレージシステム600内の全CHN110で共通にアクセス可能なLUである。

#### 【0078】

上述のように、フェイルオーバーのための引き継ぎデータにはNFSユーザデータのようにCHN110毎に個別に作成されるデータもあるし、装置管理者データのようにストレージシステム600全体で共通なデータもある。そのため本実施の形態に係る記憶デバイス制御装置100においては、引き継ぎデータの種類に応じて、異なる方法により同期が取られる。本実施の形態におけるフェイルオーバー時に引き継がれるデータと同期の方法を示す同期方法管理テーブルを図16に示す。同期方法管理テーブルは、各CHN110のメモリ113に記憶しておくようにすることもできるし、共有メモリ120に記憶しておくようにすることもできる。また、各CHN110のローカルLUに記憶しておくようにすることもできる。

#### 【0079】

図16に示す同期方法管理テーブルは、“制御情報欄”、“データの同期種別欄”、“同期方法欄”、“同期データ欄”、“ローカルLUへの反映要否欄”を含んで構成される。

“制御情報欄”には、引き継ぎデータの種類の記載される。上述したように本実施の形態においては、NFSユーザデータ、CIFSユーザデータ、装置管理者データ、フェイルオーバー用ハートビート、CHN(110)のIPアドレス、NFSファイルロック情報、クラスタ制御情報が記載される。なお以下、引き継ぎデータのことを制御情報とも記す。

#### 【0080】

“データの同期種別欄”には、各引き継ぎデータの同期を取る範囲が記載される。「クラスタ内で同期」と記載されている場合は、クラスタ内で当該引き継ぎデータの同期が取られることを示す。すなわち、更新された引き継ぎデータはクラスタ内の他のCHN110との間で同期が取られる。「ストレージ装置で同期」と記載されている場合には、ストレージシステム600全体で当該引き継ぎデータの同期が取られることを示す。「システム固有」と記載されている場合は、当



該引き継ぎデータは更新されることがないため同期を取る必要がないことを示す

”同期方法欄”には、各引き継ぎデータの同期をとる方法が記載される。「ネットワーク」と記載されている場合は、あるCHN110により更新された引き継ぎデータは、CHN110間を相互に接続するネットワークを介して他のCHN110に送信される。ネットワークとしてはLAN400を用いるようにすることもできるし、接続部150を用いるようにすることもできる。また内部LAN151を用いるようにすることもできる。「共有LU」と記載されている場合は、あるCHN110により更新された引き継ぎデータは、共有LUに記憶される。「共有メモリ」と記載されている場合は、あるCHN110により更新された引き継ぎデータは、共有メモリに記憶される。「－」と記載されている場合は、同期不要であることを示す。本実施の形態においては、装置管理者データ及びCHNのIPアドレスは更新されることがないので同期不要としているが、同期をとるようにすることも可能である。

#### 【0081】

”同期データ欄”には、各引き継ぎデータに関するコメント等が記載される。例えば引き継ぎデータの具体的なファイル名等を記載しておくことができる。同期方法管理テーブルにおいて”同期データ欄”を設けない構成とすることもできる。

#### 【0082】

”ローカルLUへの反映要否欄”は、更新された引き継ぎデータを共有LUや共有メモリ120に記載することにより同期を取った場合に、当該共有LUや共有メモリ120を共通にアクセス可能な他のCHN110のローカルLUにも当該引き継ぎデータを書き込むようにするか否かを選択するための欄である。「否」と記載されている場合は、更新された引き継ぎデータは他のCHN110のローカルLUには書き込まれない。従ってこの場合他のCHN110は、共有LUあるいは共有メモリ120にアクセスすることにより当該引き継ぎデータを参照する。「要」と記載されている場合は、更新された引き継ぎデータは他のCHN110のローカルLUにも書き込まれる。従ってこの場合、他のCHN110は自己のローカルLUにアクセスすることにより当該引き継ぎデータを参照することができる。

**【0083】**

例えば、更新頻度は少ないが参照頻度が多い引き継ぎデータは、共有LUや共有メモリ120の他、ローカルLUにも記憶するようにしておく。これにより、共有LUや共有メモリ120へのアクセス頻度を減少させ、他のCHN110との間で発生するアクセス競合を減少させることができるので、データアクセス性能を向上させることができる。また、一時的にしか参照されない引き継ぎデータや頻繁に更新される引き継ぎデータなどはローカルLUに反映させないようにする。これにより引き継ぎデータをローカルLUへ反映させる際の処理オーバーヘッドを小さくすることができる。

このように、本実施の形態に係る記憶デバイス制御装置100においては、フェイルオーバの引き継ぎデータの種類に応じた特性を考慮して、最適な方法により同期を取ることができる。

**【0084】**

次に、本実施の形態に係るフェイルオーバ時に引き継がれるデータの同期の方法を決定するための処理を示すフローチャートを図17に示す。なお以下の処理は、本実施の形態に係る各種の動作を行うためのコードから構成される障害管理プログラム705を、CPU112が実行することにより実現される。

**【0085】**

まず制御情報が発生する(S1000)。ここで制御情報が発生するとはCHN110内の他のプログラムや管理端末160内のプログラム、あるいは情報処理装置200内のプログラムから、フェイルオーバのための引き継ぎデータの更新要求を受領したという意味である。例えばファイルアクセスサービスの提供を受けるNFSユーザの追加や削除を行うために更新されたNFSユーザデータを情報処理装置200から受信した場合や、クラスタ180内のCHN110により定期的に更新されるハートビートマークの更新要求を受領した場合である。

**【0086】**

これらの引き継ぎデータの更新は自動的に行われることもあるが、オペレータが管理端末160を操作して手動で行われることもある。またユーザがNASマネージャ706により提供される設定Webページを利用することにより、情報

処理装置 200 から手動で行われることもある。自動的に行われる場合としては例えばハートビートマークが更新される場合があり、手動で行われる場合としては NFS ユーザデータが更新される場合がある。

#### 【0087】

次に、CPU 112 は同期方法管理テーブルの”データの同期種別欄”を参照し、当該引き継ぎデータが他の CHN 110 やストレージシステム 600 全体で利用されるデータであるかを確認する (S1001)。当該引き継ぎデータが他の CHN 110 と同期を必要としないデータである場合には、自己のローカル LU に当該引き継ぎデータを書き込んで終了する (S1002)。

#### 【0088】

一方、当該引き継ぎデータが他の CHN 110 と同期を必要とするデータである場合には、同期方法管理テーブルの”データの同期種別欄”を参照することにより、クラスタ内での同期が必要なデータであるかどうかを確認する (S1003)。

#### 【0089】

クラスタ内での同期が必要ではないデータである場合には、当該データはストレージシステム 600 全体での同期が必要なデータであるので、接続部 150 を介して当該データを全体共有 LU に書き込む (S1004)。これにより、当該引き継ぎデータはストレージシステム 600 内の全ての CHN 110 から同一の内容で参照されるようにすることができる。そして同期方法管理テーブルの”ローカル LU への反映要否欄”を参照することにより、当該引き継ぎデータをローカル LU にも反映するかどうかを確認する (S1005)。同期方法管理テーブルの”ローカル LU への反映要否欄”が「否」の場合はそのまま処理を終了する。”ローカル LU への反映要否欄”が「要」の場合は、S1004 において全体共有 LU に書き込んだ引き継ぎデータを、他の CHN 110 のローカル LU にも書き込む。この場合にはストレージシステム 600 内の全 CHN 110 の各ローカル LU に引き継ぎデータを書き込む (S1006)。これにより各 CHN 110 は、引き継ぎデータを参照する際には自己のローカル LU にアクセスすればよく、全体共有 LU にアクセスする必要がなくなる。全体共有 LU にアクセスする必要がなくなることにより他の CHN 110 とのアクセス競合の頻度を減少させることができるので、デ

ータアクセス性能を向上させることができる。

#### 【0090】

一方、S1003においてクラスタ内での同期が必要なデータであることが判明した場合には、クラスタ構成のCHN110を調査して、通知先CHN110を特定する(S1007)。すなわち自己と同一のクラスタに属する他のCHN110を特定する。特定は例えばCHN110の識別子により行うことができる。続いて、引き継ぎデータの同期方法にネットワークを利用するかどうかを同期方法管理テーブルの”同期方法欄”を参照することにより確認する(S1008)。”同期方法欄”の記載が「ネットワーク」以外の場合は、共有LUを使用して同期するデータであるかどうかを確認する(S1011)。同期方法管理テーブルの”同期方法欄”の記載が「共有LU」の場合には”Y”に進み、クラスタ内の共有LUに引き継ぎデータを書き込む(S1012)。これにより、当該引き継ぎデータは自己と同一クラスタ180内の全てのCHN110から同一の内容で参照されるようにすることができる。

#### 【0091】

そして必要に応じて、クラスタ内の他のCHN110に対して上記引き継ぎデータの共有LU内での記憶位置を通知するようにする(S1013)。すなわち、当該引き継ぎデータのメタデータをクラスタ内の他のCHN110に対して通知する。これにより、クラスタ内の他のCHN110は当該引き継ぎデータを参照する際に、共有LU内のどこに引き継ぎデータが格納されているのかを知ることができる。なお引き継ぎデータによっては記憶位置が特定の場所に固定されているものもある。このようなデータの場合には、引き継ぎデータの共有LUにおける記憶位置を他のCHN110に通知する必要はない。

#### 【0092】

次に同期方法管理テーブルの”ローカルLUへの反映要否欄”を参照することにより、当該引き継ぎデータをローカルLUにも反映するかどうかを確認する(S1014)。同期方法管理テーブルの”ローカルLUへの反映要否欄”が「否」の場合はそのまま処理を終了する。”ローカルLUへの反映要否欄”が「要」の場合は、S1012において共有LUに書き込んだ引き継ぎデータを、クラスタ内の他のCH

N110のローカルLUにも書き込む(S1015)。これによりクラスタ内の各CHN110は、引き継ぎデータを参照する際には自己のローカルLUにアクセスすればよく、共有LUにアクセスする必要がなくなる。共有LUにアクセスする必要がなくなることにより他のCHN110とのアクセス競合の頻度を減少させることができるので、データアクセス性能を向上させることができる。

#### 【0093】

一方S1011において、同期方法管理テーブルの”同期方法欄”の記載が「共有メモリ」の場合には”N”に進み、共有メモリ120に引き継ぎデータを書き込む(S1016)。これにより、当該引き継ぎデータは自己と同一クラスタ180内の全てのCHN110から同一の内容で参照されるようにすることができる。

#### 【0094】

そして必要に応じて、クラスタ内の他のCHN110に対して上記引き継ぎデータの共有メモリ120内での記憶位置を通知するようにする(S1017)。すなわち、当該引き継ぎデータのメタデータをクラスタ内の他のCHN110に対して通知する。これにより、クラスタ内の他のCHN110は当該引き継ぎデータを参照する際に、共有メモリ120内のどこに引き継ぎデータが格納されているのかを知ることができる。また、引き継ぎデータによっては記憶位置が特定のアドレスに固定されているものもある。このようなデータの場合には、引き継ぎデータの共有メモリ120における記憶位置を他のCHN110に通知する必要はない。

#### 【0095】

次に同期方法管理テーブルの”ローカルLUへの反映要否欄”を参照することにより、当該引き継ぎデータをローカルLUにも反映するかどうかを確認する(S1018)。同期方法管理テーブルの”ローカルLUへの反映要否欄”が「否」の場合はそのまま処理を終了する。”ローカルLUへの反映要否欄”が「要」の場合は、S1016において共有メモリ120に書き込んだ引き継ぎデータを、クラスタ内の他のCHN110のローカルLUにも書き込む(S1015)。これによりクラスタ内の各CHN110は、引き継ぎデータを参照する際には自己のローカルLUにアクセスすればよく、共有メモリ120にアクセスする必要がなくなる。共有メ

メモリ 120 にアクセスする必要がなくなることにより他の CHN 110 とのアクセス競合の頻度を減少させることができるので、データアクセス性能を向上させることができる。

#### 【0096】

一方 S1008 において、同期方法管理テーブルの”同期方法欄”の記載が「ネットワーク」の場合は”Y”にすすむ。まず引き継ぎデータを自己のローカル LU に書き込む (S1009)。そして、自己のローカル LU に書き込んだのと同じ引き継ぎデータをクラスタ内の他の CHN 110 に対してネットワークを介して送信する (S1010)。これにより当該引き継ぎデータがクラスタ内の他の CHN 110 のローカル LU にも反映され、自己のローカル LU に記憶されている引き継ぎデータの内容と、クラスタ内の他の CHN 110 のローカル LU に記憶されている引き継ぎデータの内容とを同一にすることができる。

#### 【0097】

このように、本実施の形態に係る記憶デバイス制御装置 100 においては、フェイルオーバの引き継ぎデータの種類の種に応じた特性を考慮して、最適な方法により同期処理を行うことができる。また引き継ぎデータの同期を取るようにすることにより、CHN 110 に障害が発生してからデータの引き継ぎを行う必要がなくなり、フェイルオーバを迅速に行うことが可能となる。

#### 【0098】

次に上記引き継ぎデータを参照する際の処理について説明する。引き継ぎデータの参照の処理は、本実施の形態に係る各種の動作を行うためのコードから構成される障害管理プログラム 705 を、CPU 112 が実行することにより実現される。CPU 112 は、図 18 に示す引き継ぎデータ参照テーブルを参照することにより引き継ぎデータの格納先（参照先）を知ることができる。引き継ぎデータ参照テーブルは、各 CHN 110 のメモリ 113 に記憶しておくようにすることもできるし、共有メモリ 120 に記憶しておくようにすることもできる。また、各 CHN 110 のローカル LU に記憶しておくようにすることもできる。

#### 【0099】

図 18 に示す引き継ぎデータ参照テーブルは、”制御情報欄”、”データの格納

先欄”、“データの通知有無欄”を含んで構成される。

”制御情報欄”には、引き継ぎデータの種類が記載される。本実施の形態においては、N F S ユーザデータ、C I F S ユーザデータ、装置管理者データ、フェイルオーバー用ハートビート、N F S ファイルロック情報、クラスタ制御情報が記載される。

#### 【 0 1 0 0 】

”データの格納先欄”には、当該引き継ぎデータの格納先（参照先）が記載されている。「ローカル L U」と記載されている場合には、当該引き継ぎデータは自己のローカル L U に記憶されていることを示す。すなわち、他の C H N 1 1 0 が当該引き継ぎデータを更新した際に、当該引き継ぎデータがネットワークを介して送信されてきたか、又は、他の C H N 1 1 0 により自己のローカル L U にも当該引き継ぎデータが書き込まれたかのいずれかにより、当該引き継ぎデータは自己のローカル L U に記憶されていることを示す。「共有 L U」と記載されている場合には、当該引き継ぎデータは自己が属するクラスタ 1 8 0 内の C H N 1 1 0 で共有している共有 L U に記憶されていることを示す。「全体共有 L U」と記載されている場合には、当該引き継ぎデータはストレージシステム 6 0 0 内の全 C H N 1 1 0 で共有している全体共有 L U に記憶されていることを示す。「共有メモリ」と記載されている場合には、当該引き継ぎデータは共有メモリ 1 2 0 に記憶されていることを示す。

#### 【 0 1 0 1 】

”データの通知有無欄”には、当該引き継ぎデータを更新した他の C H N 1 1 0 から、引き継ぎデータの記憶位置の通知を受けたか否かが記載される。「あり」と記載されている場合には通知を受けたことが示され、「なし」と記載されている場合には通知を受けていないことが示される。「－」と記載されている場合は通知の有無は関係ないことが示される。N F S ユーザデータの場合は図 1 6 の同期方法管理テーブルにより、ネットワーク経由で他の C H N 1 1 0 から送信されてくる。そのため、当該 N F S ユーザデータをローカル L U に記憶するのは自 C H N 1 1 0 であり、他の C H N 1 1 0 から記憶位置の通知は行われないからである。

**【 0 1 0 2 】**

このように、本実施の形態に係る記憶デバイス制御装置 1 0 0 においては、引き継ぎデータ参照テーブルを参照することにより引き継ぎデータの格納先を知ることができる。

**【 0 1 0 3 】**

次に本実施の形態に係るフェイルオーバ時に引き継がれるデータを参照する際の処理を示すフローチャートを図 1 9 に示す。

まず引き継ぎデータの参照要求を受領する (S2000)。引き継ぎデータの参照要求は、CHN 1 1 0 内の他のプログラムや管理端末 1 6 0 内のプログラム、あるいは情報処理装置 2 0 0 内のプログラムから受領する。例えばファイルアクセスサービスの提供を受ける NFS ユーザの追加や削除を行うために情報処理装置 2 0 0 から要求を受ける場合や、クラスタ 1 8 0 内の他の CHN 1 1 0 のハートビートマークを確認するために参照要求を受ける場合がある。

**【 0 1 0 4 】**

次に引き継ぎデータ参照テーブルの”データの格納先欄”を参照することにより、当該引き継ぎデータがローカル LU に記憶されているか否かを確認する (S2001)。”データの格納先欄”に「ローカル LU」と記載されている場合には、自己のローカル LU にアクセスして当該引き継ぎデータを参照する (S2002)。引き継ぎデータの記憶位置はメタデータを参照することにより知ることができる。

**【 0 1 0 5 】**

”データの格納先欄”に「ローカル LU」以外が記載されている場合には、引き継ぎデータは共有 LU、共有メモリ、又は全体共有 LU のいずれかに記載されている。そこでまず、引き継ぎデータ参照テーブルの”データの通知有無欄”を参照することにより、他の CHN 1 1 0 から当該引き継ぎデータに関する通知を受けているか否かを確認する (S2003)。

**【 0 1 0 6 】**

通知を受けていない場合には、当該引き継ぎデータは共有 LU、共有メモリ、あるいは全体共有 LU のいずれかの所定の記憶位置に記憶されている。そのため、CPU 1 1 2 は、定期的にこれらの所定の記憶位置を参照し、引き継ぎデータ



の更新がなされていないかを確認するようにする。なお上記の所定の記憶位置は、引き継ぎデータ参照テーブルに記録されるようにしておくこともできるし、引き継ぎデータ参照テーブルとは別のテーブルに記憶しておくようにすることもできる。

#### 【0 1 0 7】

S2004において一定時間の経過を待った後、引き継ぎデータ参照テーブルの”データの格納先欄”を参照することにより、クラスタ内の共有 L U に当該引き継ぎデータが記憶されているかどうかを確認する (S2007)。”データの格納先欄”に「共有 L U」と記憶されている場合には、共有 L U の所定の記憶位置にアクセスして当該引き継ぎデータを読み出す (S2008)。次に、引き継ぎデータ参照テーブルの”データの格納先欄”を参照することにより、全体共有 L U に当該引き継ぎデータが記憶されているかどうかを確認する (S2009)。”データの格納先欄”に「全体共有 L U」と記憶されている場合には、全体共有 L U の所定の記憶位置にアクセスして当該引き継ぎデータを読み出す (S2010)。”データの格納先欄”に「共有メモリ」と記憶されている場合には、共有メモリの所定の記憶位置にアクセスして当該引き継ぎデータを読み出す (S2011)。

#### 【0 1 0 8】

一方、S2003において他の C H N 1 1 0 から当該引き継ぎデータに関する通知を受けている場合には、その通知には当該引き継ぎデータの記憶位置が指定してあるか否かを確認する (S2005)。もし、当該引き継ぎデータの記憶位置が指定してある場合には、共有メモリ、共有 L U、あるいは全体共有 L U の指定された記憶位置から当該引き継ぎデータを読み出す (S2006)。

#### 【0 1 0 9】

もし、当該引き継ぎデータの記憶位置が指定されてない場合には、当該引き継ぎデータは共有 L U、共有メモリ、あるいは全体共有 L U のいずれかの所定の記憶位置に記憶されている。そこで引き継ぎデータ参照テーブルの”データの格納先欄”を参照することにより、クラスタ内の共有 L U に当該引き継ぎデータが記憶されているかどうかを確認する (S2007)。以下、上述の処理と同様の処理を行う。

**【0110】**

このように本実施の形態における記憶デバイス制御装置100においては、引き継ぎデータ参照テーブルを参照しながら上記の処理が行われることにより、その種類に応じてさまざまな位置に記憶されているフェイルオーバの引き継ぎデータを、適切に読み出すことができる。

**【0111】**

次に、本実施の形態に係るフェイルオーバ制御を示すフローチャートを図20に示す。フェイルオーバ制御は、各種の動作を行うためのコードから構成される障害管理プログラム705をCHN110が備えるCPU112が実行することにより実現される。

**【0112】**

図11で示したようにフェイルオーバ制御はクラスタ180を構成するCHN110間で行われる。フェイルオーバ制御はCHN110に障害が発生した場合の他、NASマネージャ706からの指示（フェイルオーバの実行要求）によっても行われるが、図20においては、CHN1（110）とCHN2（110）とで構成されるクラスタ内において、CHN1（110）で異常が発生した場合のフェイルオーバ制御の例を示す。

**【0113】**

まずNFS／CIFSファイル共有データがユーザによって追加される（S300）。NFS／CIFSファイル共有データとは、UNIX（登録商標）系オペレーティングシステムが実行される情報処理装置200、又はWindows（登録商標）系オペレーティングシステムが実行される情報処理装置200によりLAN400を介してアクセスされるデータをいう。追加されるとは、CHN1（110）によりNFS／CIFSファイル共有データがLUに新たに書き込まれることをいう。このとき当該NFS／CIFSファイル共有データに対応するメタデータもLUに書き込まれる。またNFSファイルロック情報も更新される。

**【0114】**

次にCHN1（110）は、NFSファイルロック情報の同期処理を行う（S3

001)。図16に示した同期方法管理テーブルの”同期方法欄”に記載されているように、NFSファイルロック情報は共有LUに記憶される。そのためCHN1(110)は共有LUに記憶されているNFSファイルロック情報を更新する。なお、同期方法管理テーブルの”ローカルLUへの反映要否欄”に記載されているように、クラスタ内の他のCHN(110)のローカルLUへの反映は行われない。

#### 【0115】

続いてS3002において、CHN2(110)は上記更新されたNFSファイルロック情報を確認する。必要に応じて自己のローカルLUに反映するようにすることもできる。

CHN2(110)は、CHN1(110)によるハートビートマークを確認し、一定時間を経過しているにも拘わらず未更新である場合や、ハートビートマークに障害発生を示す符号が記載されているものを発見した場合には、フェイルオーバ処理を開始する(S3003)。ハートビートマークは、CHN1(110)及びCHN2(110)の双方が共有メモリ120に書き込むことにより、お互いの動作状態をチェックするためのデータである。

#### 【0116】

S3004においては、CHN2(110)は引き継ぎデータ参照テーブルの”データの格納先欄”を参照することにより、NFSファイルロック情報が共有LUに記憶されていることを知ることができる。そして引き継ぎデータ参照テーブルの”データ通知有無欄”を参照し、通知がないことを知ることができる。通知がないのでNFSファイルロック情報は共有LUの所定の記憶位置に記憶されていることがわかり、CHN2(110)は共有LUの当該所定の記憶位置からNFSファイルロック情報を読み出すことができる。このようにしてCHN2(110)はCHN1(110)からNFSファイルロック情報を引き継ぐことができる。他の引き継ぎデータについても同様に、引き継ぎデータ参照テーブルを参照することによりCHN1(110)からCHN2(110)に引き継ぐことができる。これによりCHN2(110)は、それまでCHN1(110)により行われていた、情報処理装置200に対するファイルアクセスサービスを引き継いで行

うことができるようになり、フェイルオーバーが完了する (S3004)。

#### 【0117】

このように、本実施の形態に係る記憶デバイス制御装置 100 においては、引き継ぎデータの同期を取るようにすることにより、CHN 110 に障害が発生してから煩雑なデータ引き継ぎ処理を行う必要がなくなり、フェイルオーバーを迅速に行うことが可能となる。またフェイルオーバーの引き継ぎデータの種類に応じた特性を考慮して、最適な方法により同期処理を行うことができる。例えば、クラスタ内の CHN 110 でのみ同期が必要な引き継ぎデータについては共有 LU に書き込み、ストレージシステム 600 全体の CHN 110 で同期が必要なデータについては全体共有 LU に書き込む。また共有 LU に書き込んだ引き継ぎデータを、他の CHN 110 のローカル LU にも書き込むようにすることができる。これにより、各 CHN 110 は引き継ぎデータを参照する際には自己のローカル LU にアクセスすればよく、共有 LU にアクセスする必要がなくなるので、他の CHN 110 とのアクセス競合の頻度を減少させることができ、データアクセス性能を向上させることができる。

#### 【0118】

また本実施の形態における記憶デバイス制御装置 100 においては、引き継ぎデータ参照テーブルを参照しながら引き継ぎデータの参照が行われることにより、引き継ぎデータの種類に応じてさまざまな位置に記憶されているフェイルオーバーの引き継ぎデータを、適切に読み出すことが可能となる。

#### 【0119】

以上本実施の形態について説明したが、上記実施例は本発明の理解を容易にするためのものであり、本発明を限定して解釈するためのものではない。本発明はその趣旨を逸脱することなく変更、改良され得ると共に、本発明にはその等価物も含まれる。

#### 【0120】

##### 【発明の効果】

記憶デバイス制御装置、及びプログラムを提供することができる。

##### 【図面の簡単な説明】

【図 1】 本実施の形態に係るストレージシステムの全体構成を示すブロック図である。

【図 2】 本実施の形態に係る管理端末の構成を示すブロック図である。

【図 3】 本実施の形態に係る物理ディスク管理テーブルを示す図である。

【図 4】 本実施の形態に係る L U 管理テーブルを示す図である。

【図 5】 本実施の形態に係るストレージシステムの外観構成を示す図である。

【図 6】 本実施の形態に係る記憶デバイス制御装置の外観構成を示す図である。

【図 7】 本実施の形態に係るチャネル制御部を示す図である。

【図 8】 本実施の形態に係るメモリに記憶されるデータの内容を説明するための図である。

【図 9】 本実施の形態に係るディスク制御部を示す図である。

【図 1 0】 本実施の形態に係るソフトウェア構成図である。

【図 1 1】 本実施の形態に係るチャネル制御部においてクラスタが構成されている様子を示す図である。

【図 1 2】 本実施の形態に係るメタデータを示す図である。

【図 1 3】 本実施の形態に係るロックテーブルを示す図である。

【図 1 4】 本実施の形態に係るストレージシステムにおけるシステム L U、ユーザ L U、共有 L U を示す図である。

【図 1 5】 本実施の形態に係るストレージシステムにおけるローカル L U、共有 L U、全体共有 L U を示す図である。

【図 1 6】 本実施の形態に係るフェイルオーバー時に引き継がれるデータと同期の方法を示すテーブルである。

【図 1 7】 本実施の形態に係るフェイルオーバー時に引き継がれるデータの同期の方法を決定するための処理を示すフローチャートである。

【図 1 8】 本実施の形態に係るフェイルオーバー時に引き継がれるデータの参照先決定するためのテーブルである。

【図 1 9】 本実施の形態に係るフェイルオーバー時に引き継がれるデータの参

照先を決定するための処理を示すフローチャートである。

【図 20】 本実施の形態に係るフェイルオーバ処理を示すフローチャートである。

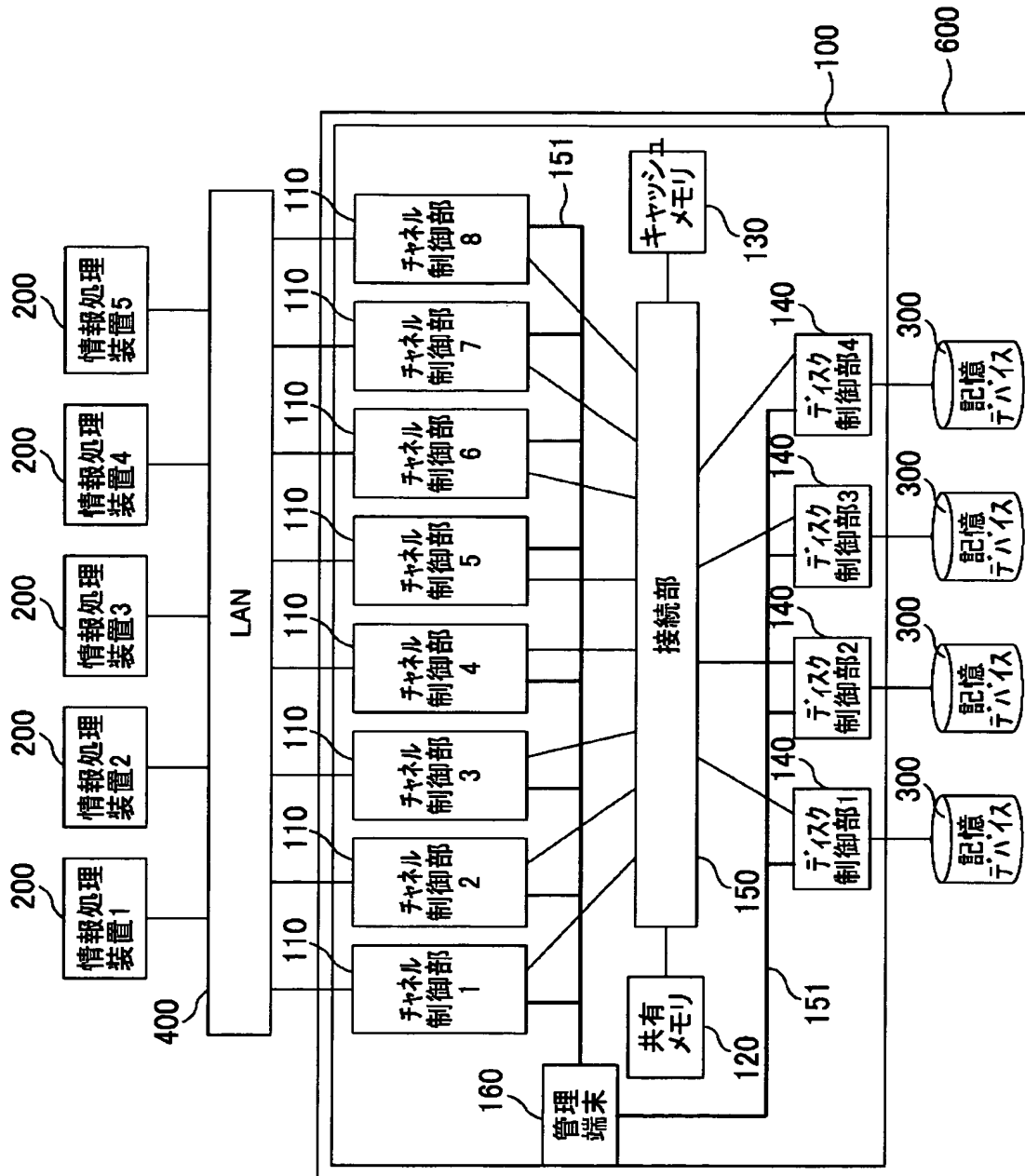
【符号の説明】

100	記憶デバイス制御装置
110	チャネル制御部
111	ネットワークインタフェース部
112	CPU
113	メモリ
114	入出力制御部
115	NVRAM
116	ボード接続用コネクタ
117	通信コネクタ
118	回路基板
119	I/Oプロセッサ
120	共有メモリ
130	キャッシュメモリ
140	ディスク制御部
150	接続部
151	内部LAN
160	管理端末
170	ファン
180	クラスタ
200	情報処理装置
300	記憶デバイス
400	LAN
600	ストレージシステム
730	メタデータ
721	ファイルロックテーブル

7 2 2      LU ロック テーブル

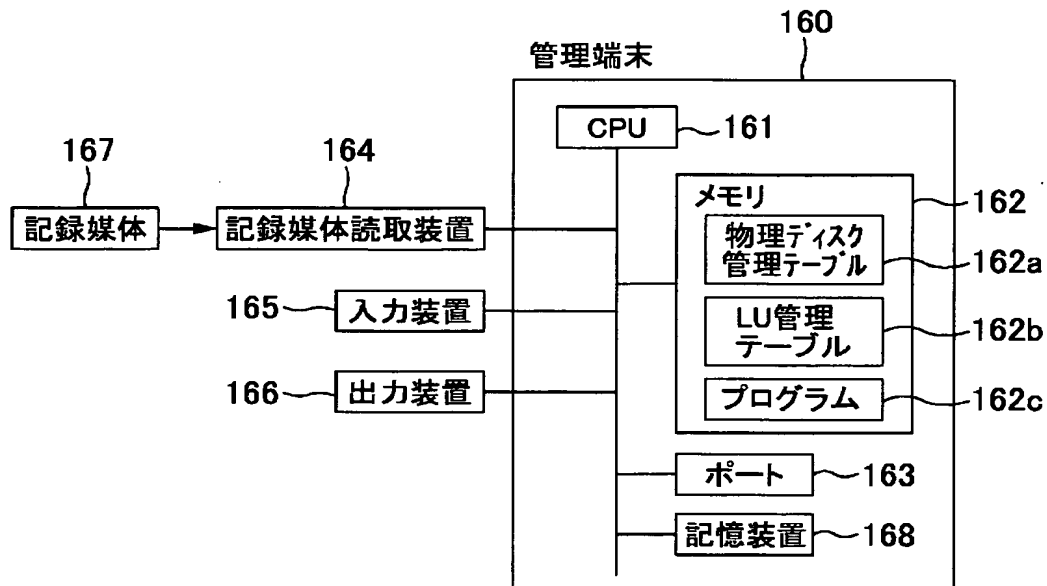
【書類名】 図面

【図 1】





【図 2】



【図 3】

162a 物理ディスク管理テーブル

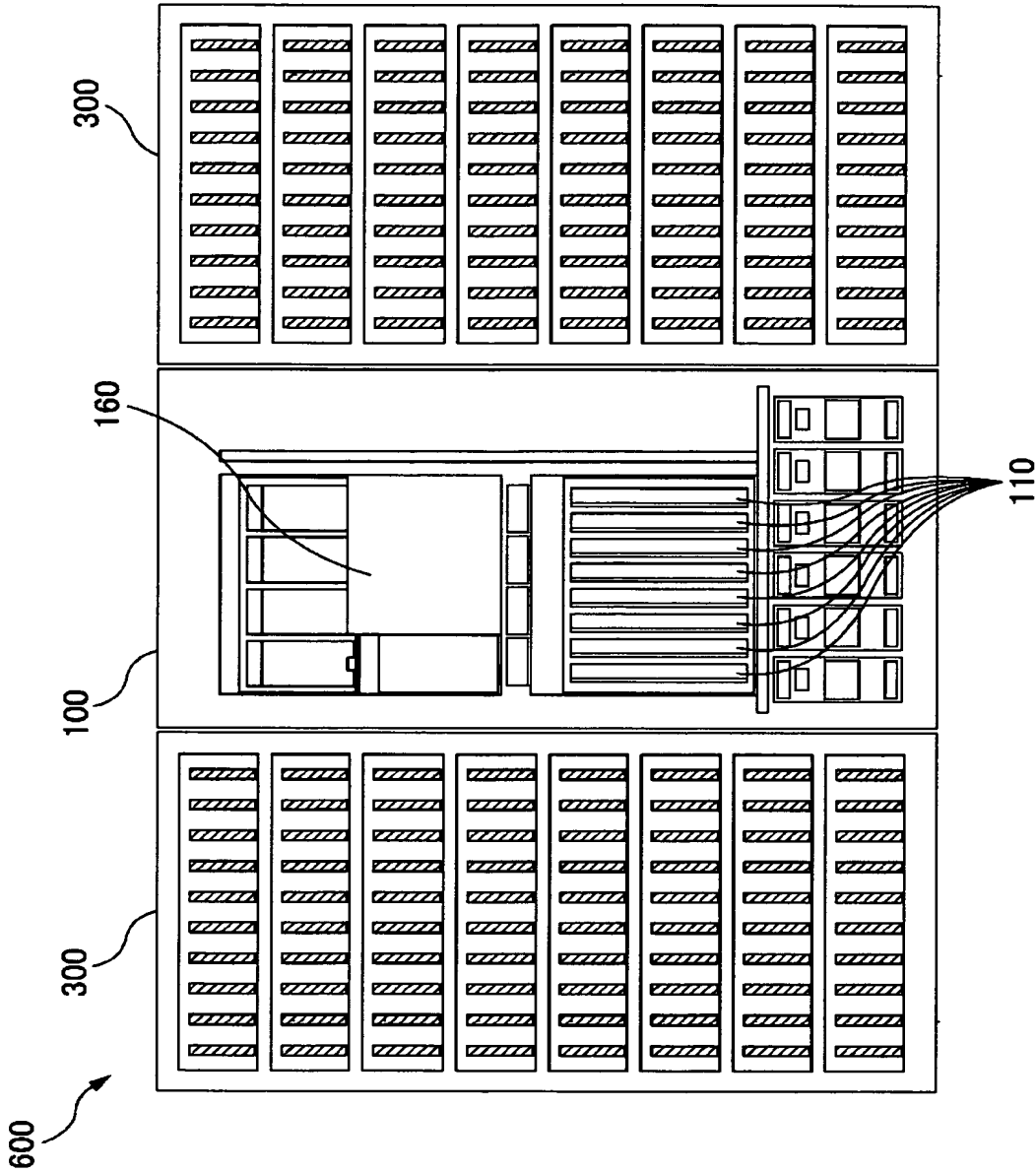
ディスク番号	容量	RAID	使用状況
#001	100GB	5	使用中
#002	100GB	5	使用中
#003	100GB	5	使用中
#004	100GB	5	使用中
#005	100GB	5	使用中
#006	50GB	—	未使用
⋮	⋮	⋮	⋮

【図 4】

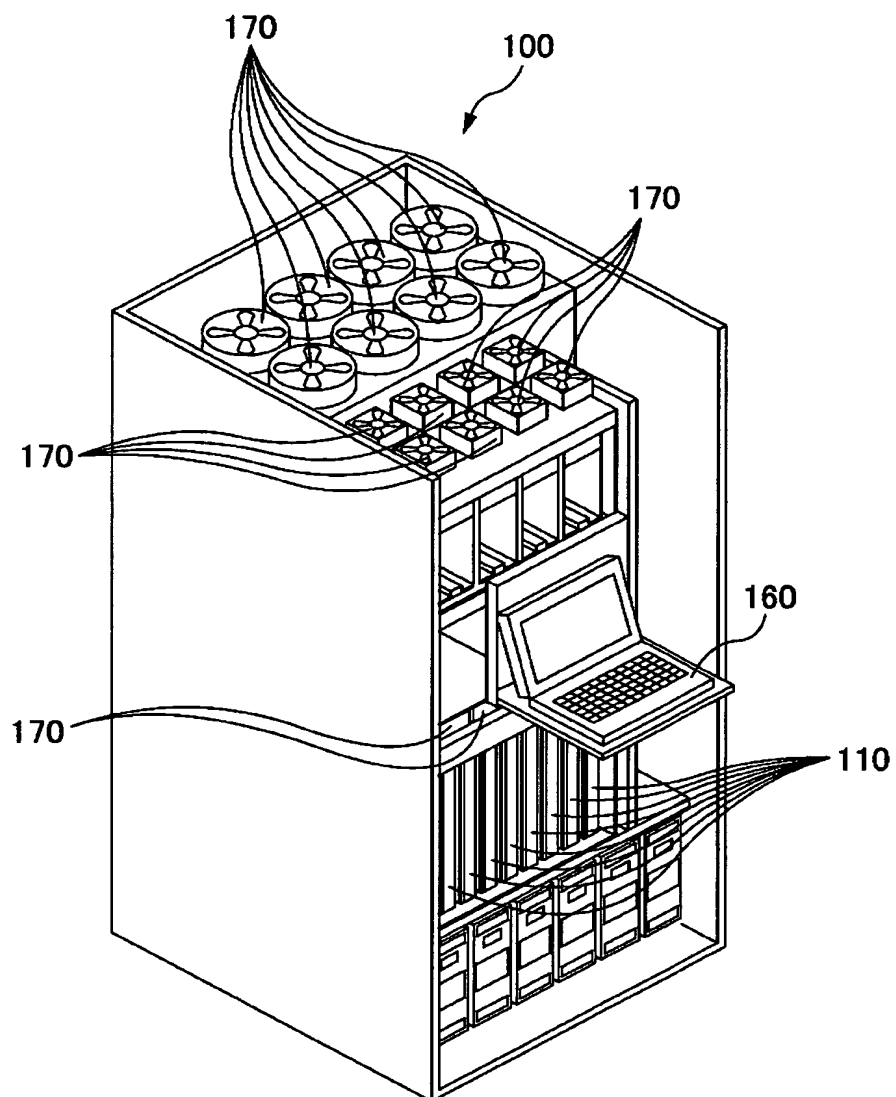
162b LU管理テーブル

LU番号	物理ディスク	容量	RAID
#1	#001,#002,#003,#004,#005	100GB	5
#2	#001,#002,#003,#004,#005	300GB	5
#3	#006,#007,	200GB	1
⋮	⋮	⋮	⋮

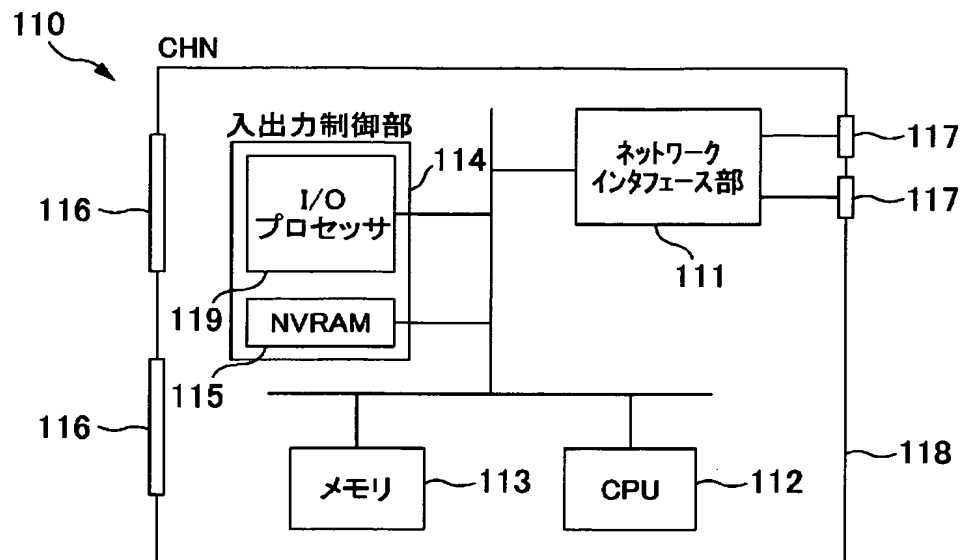
【図 5】



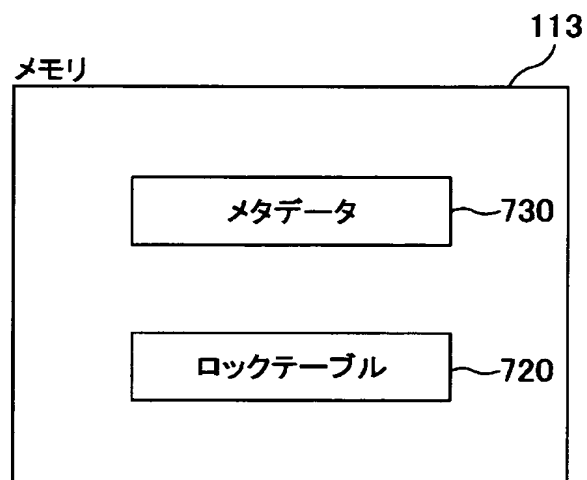
【図 6】



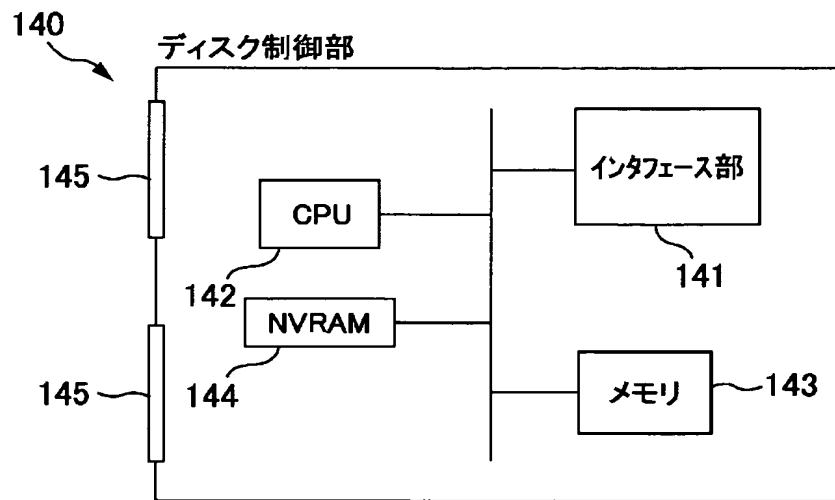
【図 7】



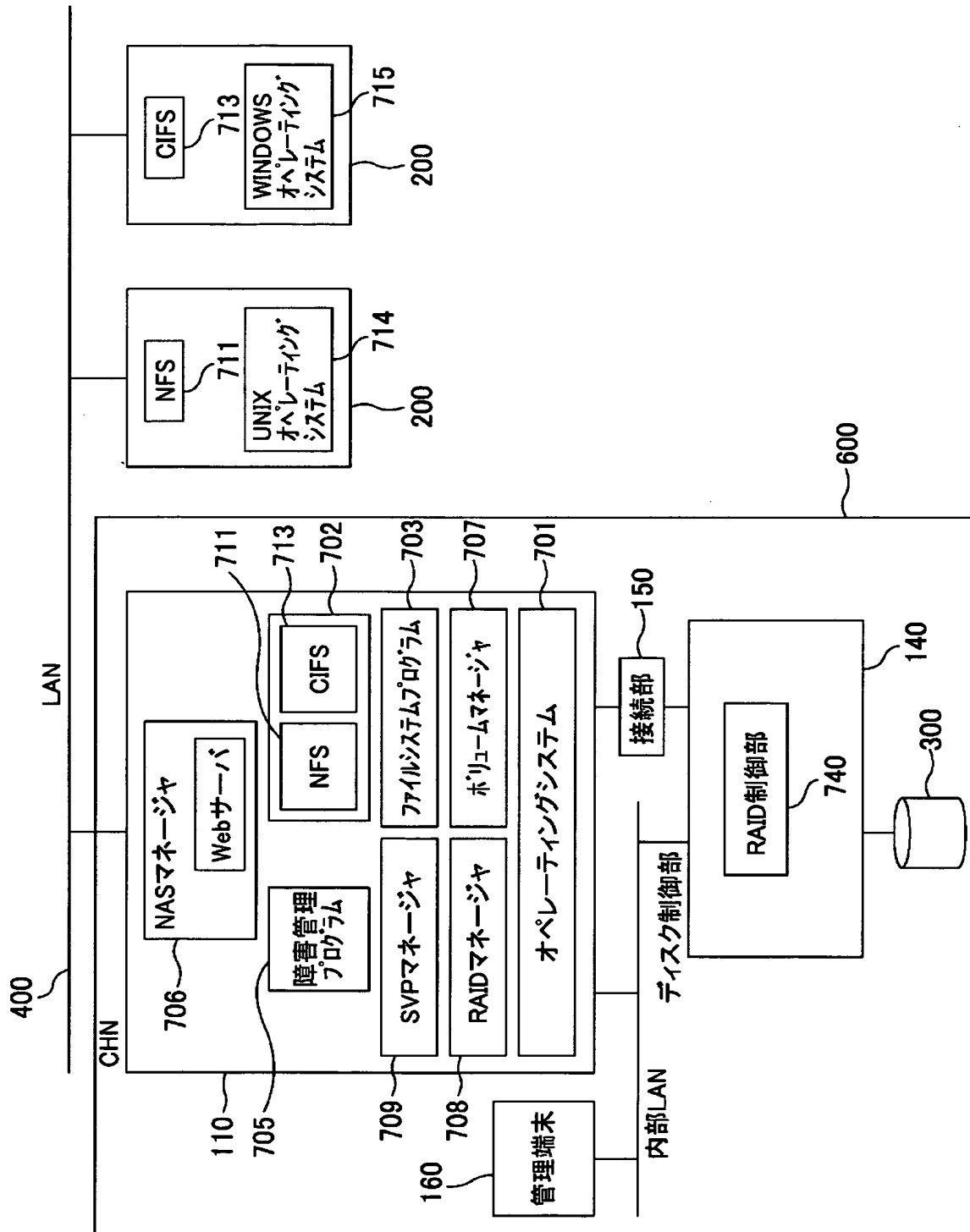
【図 8】



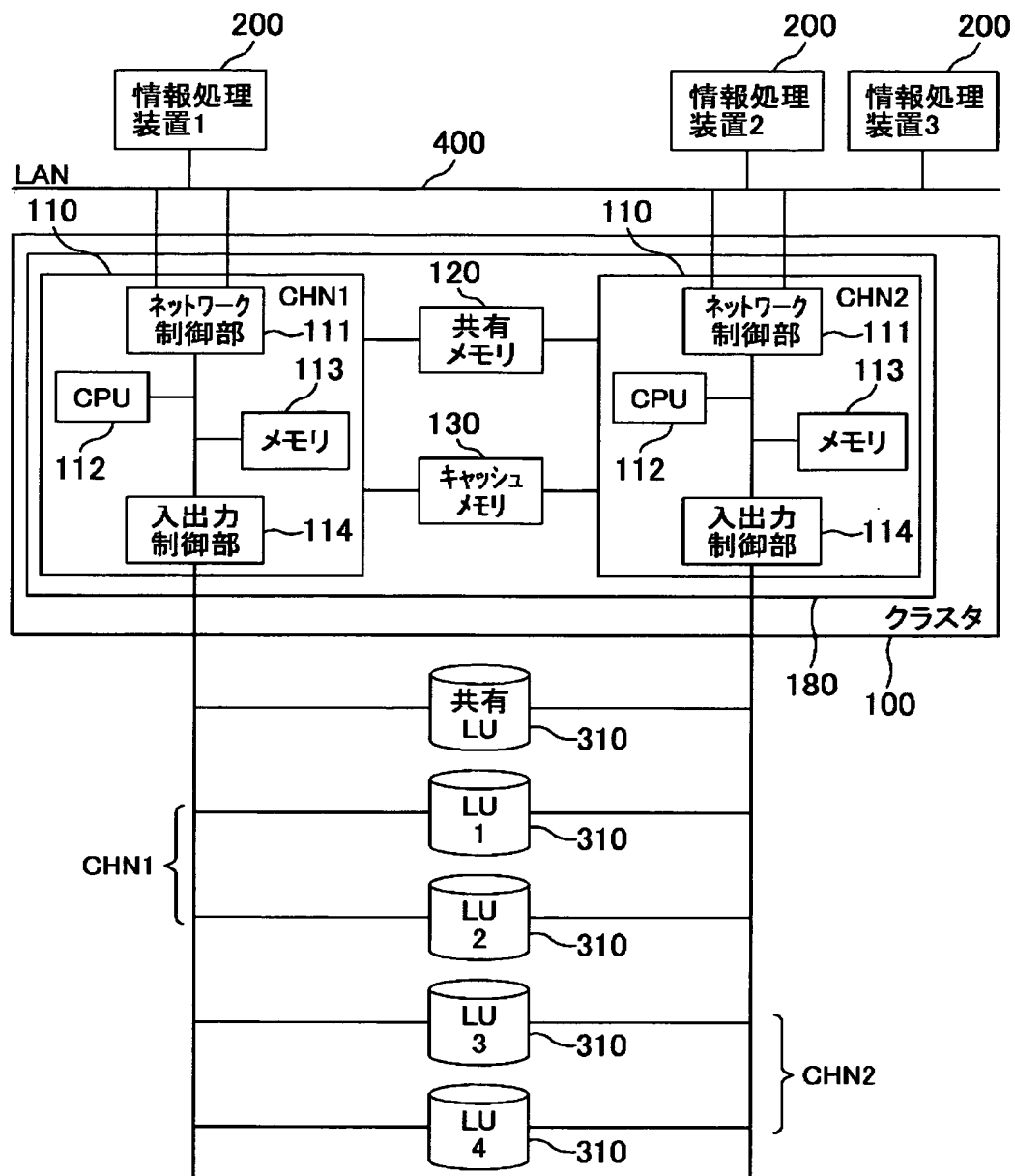
【図 9】



【図 10】



【図 11】



【図 12】

730

メタデータ

ファイル名	先頭アドレス	容量	所有者	更新時刻
A	7BSA	200MB	X	0:00
B	05BF	50MB	X	7:57
C	1F30	100MB	Y	9:15
D	470B	100MB	Z	15:20
⋮	⋮	⋮	⋮	⋮

【図 13】

721

ファイルロックテーブル

ファイル名	ロック状態
A	ロック中
B	—
C	—
D	ロック中
⋮	⋮

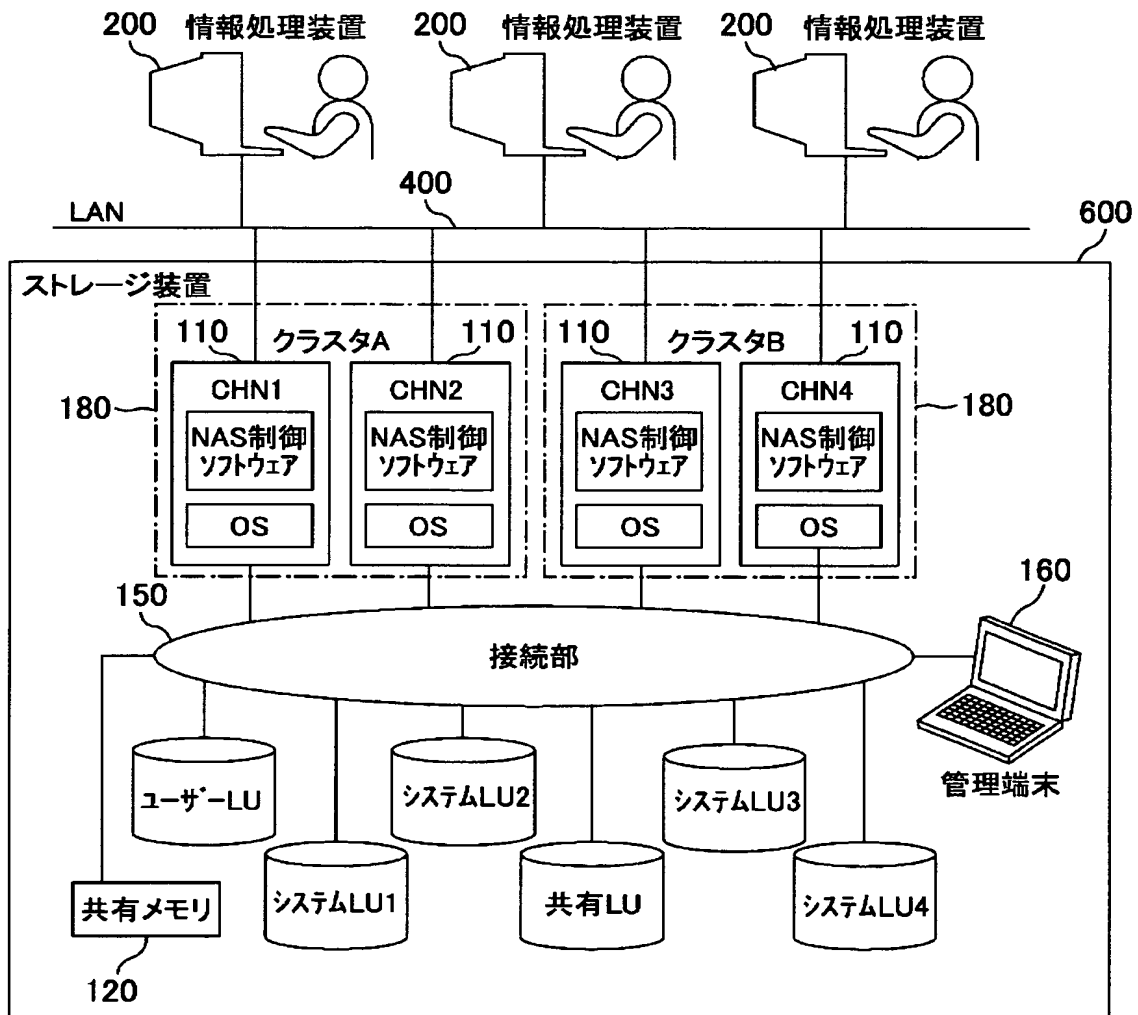
722

LUロックテーブル

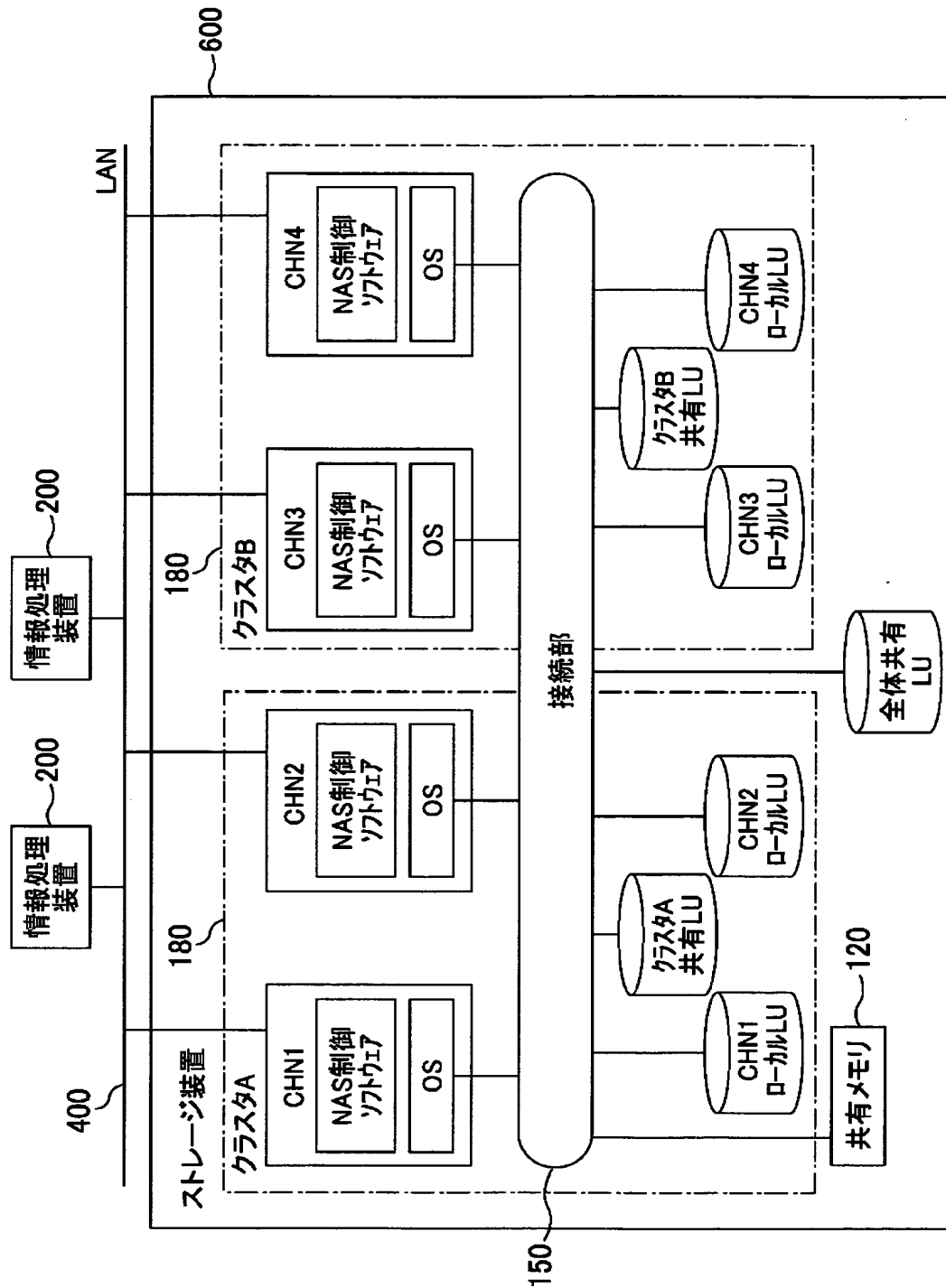
LU	ロック状態
共有	—
1	ロック中
2	—
⋮	⋮



【図 14】



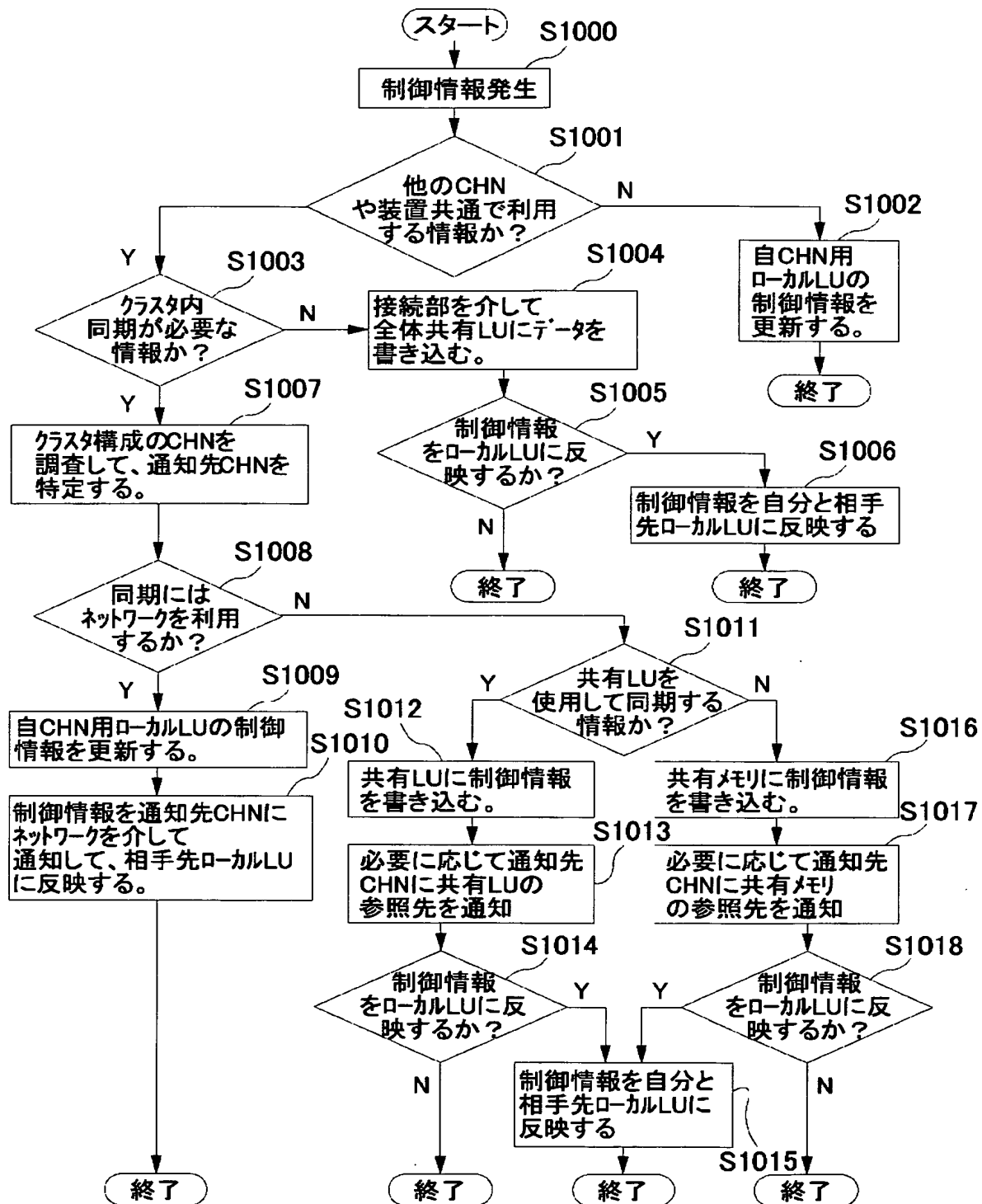
【図15】



【図 1 6】

制御情報	データの同期種別	同期方法	同期データ	ローカルLUへの 反映要否
NFSユーザデータ	クラスタ内で同期	ネットワーク	/etc/passwd	－
CIFSユーザデータ	クラスタ内で同期	共有LU	/etc/smbpasswd	否
装置管理者データ	ストレージ装置で 同期	－	－	否
フェールオーバー用 ハートビート	クラスタ内で同期	共有メモリ	ヘルスチェック (正常確認応答)	否
NAS装置固有IP アドレス	システム固有	－	/etc/network/ interfaces	－
NASファイルロック 情報	クラスタ内で同期	共有LU	ホスト名	否
クラスタ制御情報	クラスタ内で同期	共有LU	クラスタデータ ベース	要

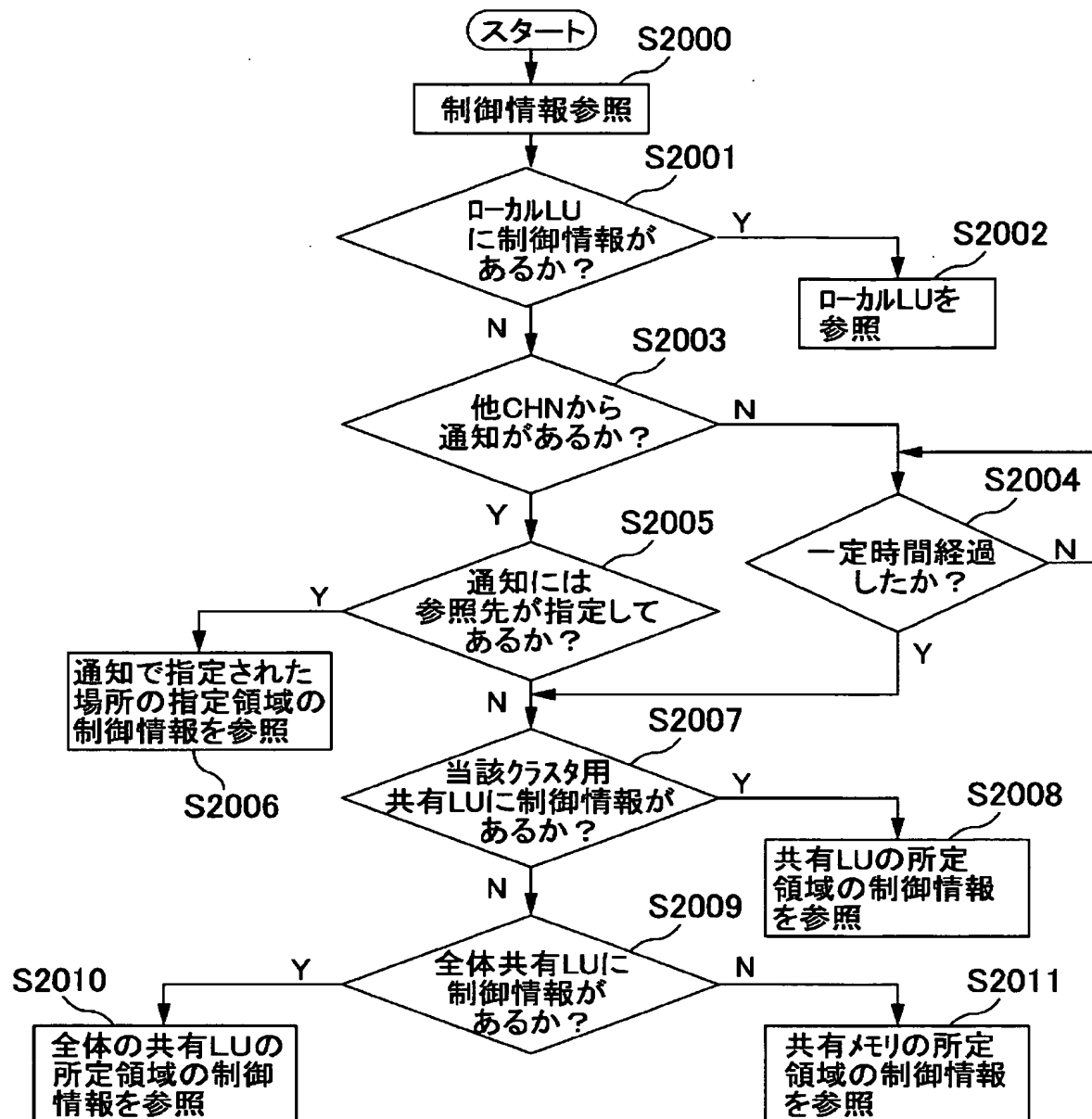
【図 17】



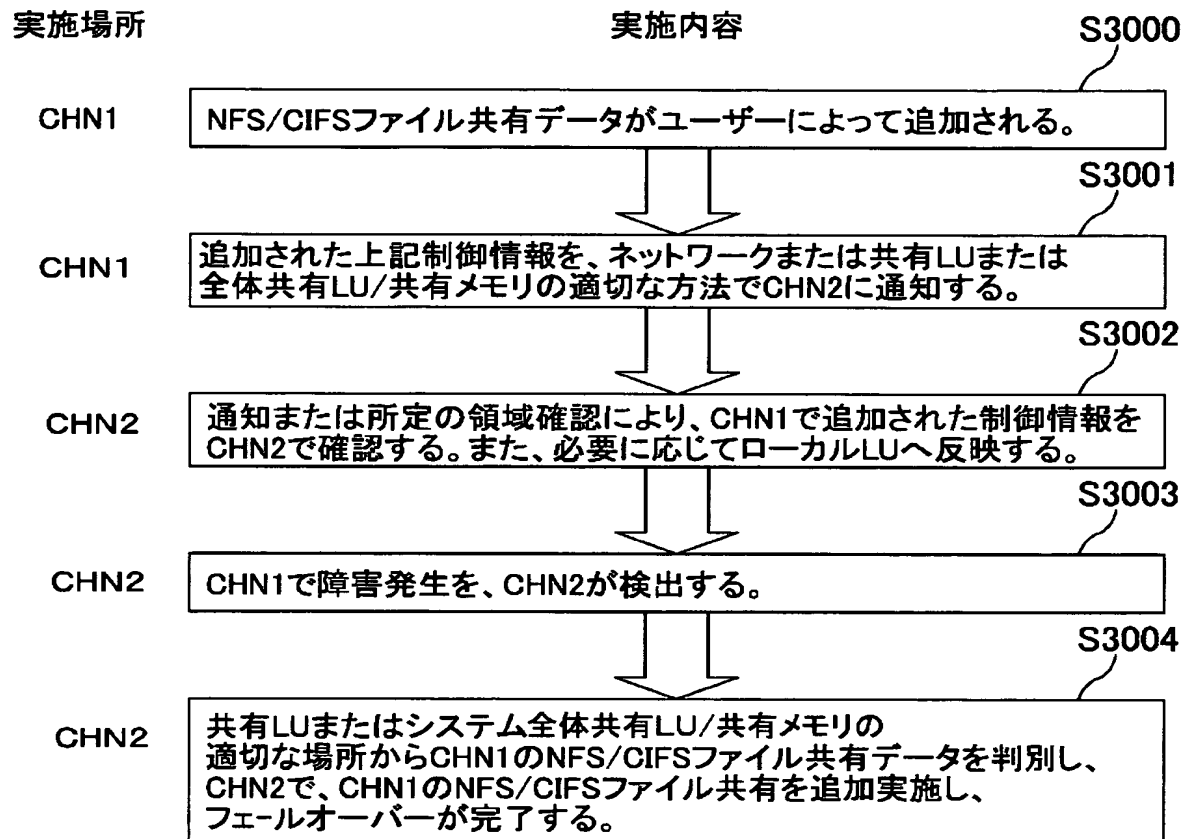
【図 1 8】

制御情報	データの格納先	データの通知有無
NFSユーザデータ	ローカルLU	—
CIFSユーザデータ	共有LU	あり
装置管理者データ	全体共有LU	なし
フェールオーバー用ハートビート	共有メモリ	なし
NFSファイルロック情報	共有LU	なし
クラスタ制御情報	ローカルLU	なし

【図 19】



【図 20】



【書類名】 要約書

【要約】

【解決手段】 情報処理装置から送信されるファイル単位でのデータ入出力要求を受信するファイルアクセス処理部と、記憶デバイスに対するデータ入出力要求に対応する I/O 要求を出力する I/O プロセッサとが形成された回路基板を有する複数のチャネル制御部を備え、各チャネル制御部はフェイルオーバーを行うためのグループに類別されている記憶デバイス制御装置であって、各チャネル制御部により更新されるフェイルオーバー時に引き継がれるデータを、当該チャネル制御部と同一のグループに類別されている各チャネル制御部が共通にアクセス可能な記憶デバイスにより提供される物理的な記憶領域上に論理的に設定される記憶領域である共有ボリュームに記憶する手段を備える。

【選択図】 図 1



特願 2003-025074

出願人履歴情報

識別番号

[000005108]

1. 変更年月日

1990年 8月31日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台4丁目6番地

氏 名

株式会社日立製作所